The Facilitation of Protein-DNA Search and Recognition by Multiple Modes of Binding

Thesis advisors

Leonid Mirny & Antoine van Oijen

Author

Jason Suh Leith

The Facilitation of Protein-DNA Search and Recognition by Multiple Modes of Binding

Abstract

The studies discussed in this thesis unify experimental and theoretical techniques, both established and novel, in investigating the problem of how a protein that binds specific sites on DNA translocates to, recognizes, and stably binds to its target site or sites. The thesis is organized into two parts. Part I outlines the history of the problem and the theory and experiments that have addressed the problem and presents an apparent incompatibility between efficient search and stable, specific binding. To address this problem, we elaborate a model of protein-DNA interaction in which the protein may bind DNA in either a search (**S**) mode or a recognition (**R**) mode. The former is characterized by zero or weak sequence-dependence in the binding energy, while the latter is highly sequence-dependent. The protein undergoes a random walk along the DNA in the **S** mode, and if it encounters its target site, must undergo a conformational transition into the **R** mode. The model resolves the apparent paradox, and accounts for the observed speed, specificity, and stability in protein-DNA interactions. The model shows internal agreement as regards theoretical and simulated results, as well as external agreement with experimental measurements.

Part II reports on research that has tested the applicability of the two-mode model to the tumor suppressor transcription factor p53. It describes in greater depth the experimental techinques and findings up to the present work, and introduces the techinques and biological system used in our research. We employ single-molecule optical microscopy in

two projects to study the diffusional kinetics of p53 on DNA. The first project measures the diffusion coefficient of p53 and determines that the protein satisfies a number of requirements for the validity of the two-mode model and for efficient target localization. The second project examines the sequence-dependence in p53's sliding kinetics, and explicitly models the energy landscape it experiences on DNA and relates features of the landscape to observed local variation in diffusion coefficient. The thesis closes with proposed extensions and complements to the projects, and a discussion of the implications of our work and its relation to recent developments in the field.

Contents

List of Figures

Citations to Previously Published Work

The theory summarized in Chapter 1, including small portions of the text, appears in

> Mirny LA, Slutsky M, Wunderlich ZB, Tafvizi A, Leith J, Kosmrlj A (2009) How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. A.* **42**:434013.

Most of Chapter 4 has been previously published as

> Tafvizi A, Huang F, Leith JS, Fersht AR, Mirny LA, van Oijen AM (2008) Tumor Suppressor p53 Slides on DNA with Low Friction and High Stability. *Biophys. J.* **95**:L01-L03.

Parts of Chapter 5 appear in

> Leith JS, Tafvizi A, Huang F, Uspal WE, Doyle PS, Fersht AR, Mirny LA, van Oijen AM. Sequence-dependent sliding kinetics of p53. Submitted to *Proc. Natl. Acad. Sci. USA*.

Acknowledgments

I thank foremost my advisors, Leonid Mirny and Antoine van Oijen, for all the usual things one thanks ones advisors for—knowledge, patience, wisdom, etc.—but especially for their flexibility and compatibility as co-advisors. Neither was jealous of my time when I was working principally with them nor distant or uninterested when I was working principally with the other. Moving between projects was far smoother than I had any reason to expect, and I think that Antoine's and Leonid's complementary talents and approaches to research gave me a rare opportunity to learn from and work according to the best of both worlds. They were both dedicated to the success of my projects while at the same time leaving me the freedom to explore paths as I saw fit. In their distinct ways, the work environments they fostered were informal and intellectually dynamic—while the lion's share of the credit for learning from my colleagues must of course go to my colleauges themselves, Antoine and Leonid created atmospheres where people felt that they were all part of one big team, despite having their own often dissimilar projects. I also thank them both for their humanity and personal support that is matched by few principal investigators.

My experimental work would not have been possible without the sharing of labeled protein and advice and troubleshooting on its use, by Sir Alan Fersht, and the researchers Fang Huang and Sridharan Rajagopalan in his lab. I also wish to thank collaborators Patrick Doyle and his student William Uspal.

Early in my graduate school career, I had the chance to learn from course instructors and rotation advisors. Easily the most useful and equally easily the most difficult course I took in graduate school was an applied math course instructed by Michael Brenner, not so much for any particular techniques it taught me but more for its having provided a general intellectual approach to various problems. My rotations with Eugene Shakhnovich and with George Whitesides and his post-doctoral fellow Demetri Moustakas contributed similarly to my intellectual development.

Discussions with my dissertation advisory committe, consisting of Shamil Sunyaev, Jagesh Shah, and Martha Bulyk, were helpful in preparing me for my papers, my thesis, and my job talks. I must also thank, admittedly in advance, my thesis defense committee: Joseph Loparo, William Shih, and Jeremy England.

Especially in my experimental work, my projects would have gone nowhere without the support, both scientific and moral, of my fellow lab members. In the van Oijen lab, I give especial thanks to Anna Kochaniak, Candice Etson, and Joseph Loparo, whose willingness to take time out from their own projects to teach me techniques relevant to mine was limitless. That Anna, Candice, and Joe were so unreasonably helpful should not detract from the generosity of the other members of the van Oijen lab, Nathan Tanner, Jason Otterstrom, Hasan Yardimci, Sam Hamdan, Satoshi Habuchi, Daniel Floyd, Mark Elenko, and Peter Stark.

Theoretical work can sometimes be a lonelier pursuit than experimental work, but the members of the Mirny lab guaranteed that I always felt part of a group. Our discussions would often wander, much like a protein looking for its cognate site, but would hit upon unexpected scientific and personal insights with a remarkable frequency. My good cheer and mental acuity owes a great deal to Christopher McFarland, Maxim Imakaev, and Geoff Fudenberg. I also wish to thank Grisha Kolesov and Zeba Wunderlich for helping get my feet planted in the lab, and the more recent members, Anton Goloborodko and Jaie Woodward.

Anahita Tafvizi, my labmate in both the van Oijen and the Mirny labs, gets her own special paragraph. Anahita and I worked side by side in our experimental and data analytical work. Her experience in working with p53 and collecting and processing data from the movies was invaluable. We shared reagents, equipment, results, and code freely. After she graduated, the benificent ghost of Anahita stayed around to give me important control data, a model thesis, and helpful professional discussions.

No Harvard Biophysics student's acknowledgements could resemble being complete without expressing gratitude to and appreciation for Jim Hogle, the chair of the program, and Michele Jakoulov, the program administrator. Jim is always willing to serve in whatever official capacity is needed to make things work out, all while staying easy-going and in good humor. Michele is dedicated to the Harvard Biophysics program and to the individual students in it in a way I have never witnessed from anyone in a similar position as she is. She makes sure that all the business gets done, and also the fun, through her event planning and more generally contributing to the collegial feel of the program. I also am thankful for the financial support provided by a National Science Foundation Graduate Research Fellowship.

A single hallway at Harvard Medical School was populated by four colleague-friends whose talent, love, and spine helped me through professional and personal rough spots: Candice Etson, Anna Kochaniak, Jonathan Schneiderman, and Peter Stark.

Outside of an academic context, I'm thankful for my (formerly) Boston-based friends Jacob Eisler and Will Carspecken for keeping my life rich and fun, and facilitating my growth and strengthening as a person. I'm also thankful for my past and current easy-going, no-drama roommates, Ezra Keshet, Darrick Yee, and Chris Sanders, who helped make the environment I came home to at the end of a long day in lab a relaxing one.

For most of my time in graduate school, I had the joy of being with my partner Jonas Nahm, who kept me inspired to work hard but efficiently and who was a source of constancy amidst moving back and forth between projects and labs. He calmed me when I was distraught, provided unreasonably sound advice on all ranges of subjects, and celebrated with me when I had reason to celebrate and gave me reason to celebrate when he would celebrate. Even after our relationship ended, his words helped me through the challenges at the end of grad school.

Almost finally, I thank my boyfriend Robert Fung for amazingly generous-yet-rational love. At once forbearing and exuberant, he has made the home stretch of my graduate work a saner and happier time that it might have been. Especially during the thesis-writing phase, his support and patience, allied with that of my parents at whose home I stayed awhile, was a boon.

My mother and father, Suzette Leith and Tony Suh, have been devoted, liberal, and supportive my entire life, with the last six and a half years no exception. Their love precedes their relationship with me; it makes as much sense to ask upon what it depends as it does to ask upon what protein structures the Pythagorean theorem depends.

Part I

Two-mode model for the facilitated diffusion of proteins to target sites on DNA

Chapter 1

Introduction

Every eukaryotic cellular process, such as gene expression, signal transduction, catalysis, and DNA repair, depends on the ability of biomolecules to locate and reliably recognize each other or a particular conformation or activation state. Since the cell, however, is far from equilibrium, the specificity of biomolecules or parts of biomolecules for each other depends not only on thermodynamic but also on kinetic considerations: can two molecules bind, and can they bind each other *fast enough*?

This thesis discusses background for this problem of molecular search and recognition and presents the results of a unified experimental and theoretical approach to the problem. It focuses on how DNA-binding proteins, and in particular transcription factors, locate and recognize their target sites amidst a vast excess of non-target DNA and relying solely on passive transport. Part I presents a model of protein-DNA interaction in which the complex may occur in either a search mode or a recognition mode, and Part II discusses single-molecule experiments performed that support this model. Separate introductory chapters precede the presentation of the results and techniques for the theoretical and the experimental work, although the motivation and background for both shares a great deal.

1.1 Motivation for 1D-3D model

Molecular recognition is a major field within the biophysical community. Research in the field can largely be divided into two related questions: whether two or more biomolecules will bind, and how two or more biomolecules bind. The first is concerned with predicting or engineering binding affinity as a function of structure and sequence, while the second seeks to elucidate the ways in which molecules approach each other and undergo the necessary changes in conformation and orientation to form a complex. Both questions are asked of interactions between proteins and small molecules [1, 2, 3, 4], proteins and other proteins [5, 6, 7, 8], and proteins and DNA [9, 10, 11], and both questions admit theoretical [8, 11] and experimental [7, 10] approaches.

In most cases, biomolecules are envisioned to bind each other from solution. Of special importance to protein-DNA interactions, however, is how DNA-binding proteins (DBPs) find their specific DNA sequences prior to actually binding them. This process is not trivial, as many classes of DBPs, including transcription factors (TFs), locate their target sites using only passive transport. The rate at which transcription factors can bind their target sites is of utmost biological importance for time-sensitive processes, such as response to heat shock or DNA damage. Generally, two molecules in solution absent an active transport mechanism cannot associate with each other at a rate faster the diffusion-limited rate, which in the case of a mobile protein and a specific DNA sequence that is assumed to move much more slowly than the protein, is given by Smoluchowski as:

$$k_{\text{smol}} = 4\pi D_{3D} b a \tag{1.1}$$

where D_{3D} is the diffusion coefficient of the protein in solution, b is the linear size of the target, which for DNA can be assumed to be no greater than the spacing between base pairs, 0.34nm, and a is the fraction of collisions resulting in binding. Proteins in aqueous

environments have $D_{3D} \approx (1-5)\times 10^{-6}$ cm^{2}s^{-1}. With $a \approx 0.2-0.5$ given that electrostatic interactions between negatively-charged DNA and basic amino acids can orient a protein favorably on its approach to DNA, we obtain a diffusion-limited $k_{\mathrm{smol}} \approx 10^8 M^{-1} s^{-1}$.

This diffusion limit, however, appeared to be broken in a study by Riggs *et al.* which measured the association rate of the *lac* repressor protein to its binding site in the *lac* operon as 10^{10} M^{-1} s^{-1} [12]. This rate is nearly two orders of magnitude greater than the diffusion limit, and motivated researchers to explain how it might be possible.

One mechanism suggested to explain the greatly accelerated binding rate observed by Riggs *et al.* was one of facilatated diffusion, consisting of alternating rounds of 3D random walks by the protein in solution and 1D random walks along the DNA (Figure 1.1). Before the Riggs experiment with *lac* repressor, Adam and Delbrück [13] suggested that a reduction in dimensionality of diffusive searches in biological systems could speed target localization. Riggs *et al* [14] considered the 1D/3D mechanism unlikely, but it was taken up again by Richter and Eigen [15], and developed by Berg and Blomberg [16] and by Berg *et al* [17].

The 1D/3D mechanism requires that DBPs employing it have non-specific affinity to DNA. This affinity was understood by Riggs *et al.* in their study of *lac* repressor and informed their consideration of the mechanism, even if they ulitmately rejected it. Binding energy between DBPs and non-specific DNA has been measured for several DNA–binding proteins to have a range of 10–15 k_BT (at physiological salt concentration), owing largely to electrostatic interactions between charged moieties [18]. The non-specific binding energy is thus highly sensitive to ionic strength, as will be seen to be important in Chapter 4.

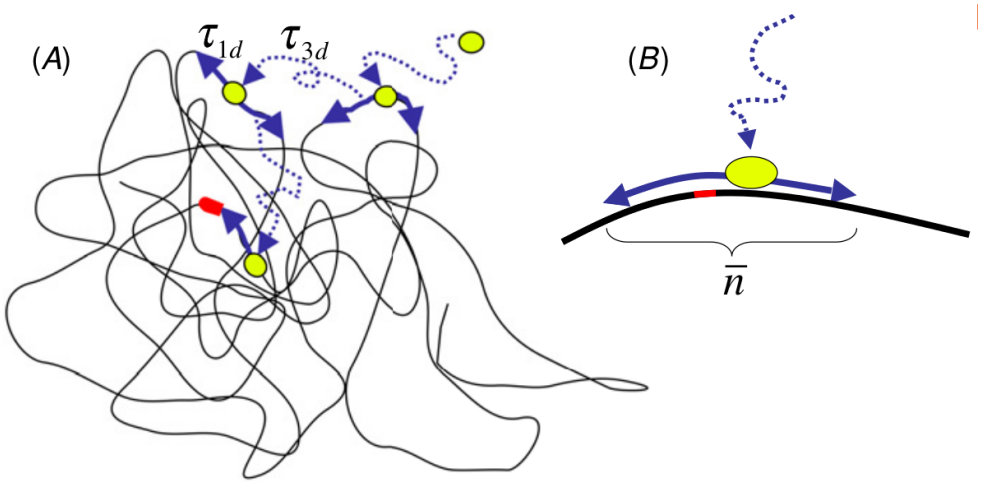


Figure 1.1: (**A**) The mechanism of facilitated diffusion. The search process consists of alternating rounds of 3D and 1D diffusion, each with average duration τ_{3D} and τ_{1D}, respectively. (**B**) The antenna effect [19]. During 1D diffusion (sliding) along DNA, a protein visits on average $\bar{n}$ sites. This allows the protein to associate some distance $\sim\bar{n}$ away from the target site and reach it by sliding, effectively increasing the reaction cross-section from 1 base-pair to $\sim\bar{n}$ bp. The antenna effect is responsible for the speed-up by facilitated diffusion.

1.2 Experimental evidence for 1D/3D model

The first directly controlled experiment supporting 1D/3D facilitated diffusion, in 1982, examined the effect of non-specific DNA flanking the restriction enzyme EcoRI's scissile site [20]. Jack *et al.* found that EcoRI cut its target site up to 8 times faster when the length of linear plasmids in which the site was located was increased from 34 to 4,000 bp. The use of alternating rounds between 1D and 3D search was strongly suggested by a study on another restriction enzyme, EcoRV [21]. The researchers measured the rates of DNA cleavage by the enzyme on three DNA constructs: a 3.4-kb circular plasmid with EcoRV's target site, the same plasmid but with the target site and another a 0.3 kb separated into a minicircle which was concatenated with the remaining 3.1 kb of the circular plasmid, and the 0.3-kb minicircle alone. Rates of DNA cleavage were indistinguishable for the single 3.4-kb plasmid and the catenane, but were reduced on the unconcatenated minicircle. The

results can be explained if the protein, after binding the 3.1-kb plasmid and undergoing 1D diffusion to the minicircle, could transfer efficiently between DNA molecules and resume a 1D search on the smaller circle.

More recently, direct single molecule evidence of site-specific DBPs translocating in 1D along non-cognate DNA has appeared for a wide range of DBPs, including transcriptions factors. The *lac* repressor itself was found by Wang *et al.* to indeed undergo 1D random walks on DNA, contrary to Riggs *et al.*'s conclusion, by fluorescently labeling the protein and imaging it translocating along DNA [22]. Their measurements of the protein's 1D diffusion coefficient, D_{1D}, and the distance along DNA it diffused between association and dissociation gave an estimate for the enhancement of the rate at which it associated to its target site above the diffusion limit as a factor of 90, in broad agreement with Riggs's bulk biochemical studies.

Another class of site-specific DBPs that have been directly visualized translocating along DNA are those involved in DNA repair [23, 24, 25]. Graneli *et al.* imaged Rad51, a protein specific for double–stranded DNA breaks, undergoing 1D random walks on λ–phage DNA, and in another study, Blainey *et al.* visualized the same for human oxoguanine glycosylase (h0gg1) [25]. Beyond individual proteins, the *E. coli* MutS–β sliding clamp complex was found similarly to slide on DNA [26].

1.3 Earlier development of our theory; motivation for two-mode model

1.3.1 One-mode model

Many groups [27, 28, 29, 30, 31, 32, 33, 34, 35, 36] have offered further elaboration based on the 1D/3D model of Berg [17]. The approach off of which this thesis's theoretical

work is based [27, 37] is intended to be transparent and intuitive. It considers a single protein searching for a single target site on a long DNA molecule of M bps by the 1D/3D mechanism. The search consists of multiple rounds, each consisting of one round of 1D diffusion followed by one round of 3D diffusion. The total search time, t_s, equals:

$$t_s = \sum_{i=1}^{k} (\tau_{1D,i} + \tau_{3D,i}), \tag{1.2}$$

where $\tau_{1D,i}$ and $\tau_{3D,i}$ are the durations of 1D and 3D diffusion in the ith round of searching, and k is the number of rounds until the target site is found. With M total positions in the genome and an average of $\bar{n} \ll M$ base–pairs scanned per 1D round, then the average total time of the search can be written as

$$\bar{t}_s = \frac{M}{\bar{n}} (\tau_{1D} + \tau_{3D}). \tag{1.3}$$

Assuming the time spent on DNA during a 1D-diffusion round is exponentially distributed, with mean τ_{1D}, the mean number of visited sites $\bar{n}$ equals [38]

$$\bar{n} = 2\sqrt{D_{1D}\tau_{1D}}. \tag{1.4}$$

where D_{1D} is the diffusion coefficient for 1D diffusion. Substituting Equation 1.4 into Equation 1.3 and setting $d\bar{t}_s/d\tau_{1D} = 0$ shows that an equal partition of the protein's time into 1D and 3D diffusion, *i.e.* $\tau_{1D} = \tau_{3D}$, yields a optimal search time

$$\bar{t}_{opt} = \frac{2M}{\bar{n}} \tau_{3D} = M\sqrt{\frac{\tau_{3D}}{D_{1D}}}. \tag{1.5}$$

Equations 1.3–1.5 may be used to determine the speed-up due to 1D/3D facilitated diffusion relative to a 3D–only or 1D–only mechanism. For 3D diffusion alone, one sets $\tau_{1D} = 0$ and $\bar{n} = 1$, yielding $\bar{t}_{3D} = M\tau_{3D}$, which is $\bar{n}/2$ times slower than what the 1D/3D mechanism achieves. The search time by 1D diffusion alone is $\bar{t}_{1D} \approx M^2/D_{1D}$, which is $M/\bar{n}$ times slower.

1.3.2 The double-edged sword of non-specific binding

These speed–ups owing to facilitated diffusion assume optimal 1D/3D partitioning $\tau_{1D} = \tau_{3D}$. The abundance of DNA in the cell and general property of site-specific DBPs to have affinity for non-specific DNA [18], however, prevent this optimum partitioning from obtaining. Applying the formalism presented above to the Smoluchowski rate for diffusion-limited binding (Equation 1.1) allows us to readily understand how non-specific binding can have a slow-down effect as well as a facilitating effect.

The rate and the mean time of the search process are connected by $\bar{t_s} = (k_s[T])^{-1}$, where $[T]$ is the concentration of the target sequence, which is related to the total DNA concentration: $[T] = [\text{DNA}]/M$. Note that τ_{3D} is the mean diffusion-limited time experienced by the protein before it interacts with *any* region of DNA, and thus, $\tau_{3D} = (k_{\text{smol}}[\text{DNA}])^{-1}$. Using these expressions and Equation 1.3 for the mean search time, we arrive at the rate of the search reaction

$$k_s \approx k_{\text{smol}}\Big(\frac{\tau_{3D}}{\tau_{1D} + \tau_{3D}}\Big)\bar{n} = 4\pi D_{3D}\Big(\frac{\tau_{3D}}{\tau_{1D} + \tau_{3D}}\Big)\bar{n}a \tag{1.6}$$

Two aspects of the search process become transparent from this equation. First, the acceleration of search by sliding effectively increases the cross-section from $b = 1$ bp to $\bar{n}$ bp of DNA, allowing the protein to reach the target site by associating with $\bar{n}$ base-pairs around it. Hu *et al.* [19] called this the *antenna effect* (Figure 1.1B). The second effect is the slow-down due to non-specific binding of the protein to DNA. While searching for its target, the protein spends a certain fraction of its time bound to DNA far from the target and, thus, not diffusing in 3D. This effect is manifested by the factor $\tau_{3D}/\tau_{1D} + \tau_{3D}$, which is the fraction of time the protein spends diffusing in 3D. Thus, binding non-specifically to DNA leads to a reduction of spatial mobility, which can be taken into account by an effective diffusion coefficient $D_{3D,eff} = D_{3D}\tau_{3D}/\tau_{1D} + \tau_{3D}$.

Importantly, the slow-down term depends upon a proteins affinity for non-specific DNA and the DNA concentration, but not upon the rate at which it slides along DNA. The speed-up term $\bar{n} = 2\sqrt{D_{1D}\tau_{1D}}$ (Equation 1.4), in contrast, depends on the absolute time spent in each round of sliding and the diffusion coefficient of sliding. Taken together the two effects can lead to speed-up (up to $\sim\bar{n}$ times) or slow-down as compared to the search by 3D diffusion alone. A similar observation that 1D/3D mechanism can lead to a slow search was made by Hu *et al.* [19].

1.3.3 Diffusion on a sequence-dependent landscape: the search-stability paradox

For facilitated diffusion to be an effective mechanism, a sliding protein must read the DNA sequence over which it is sliding, which is tantamount to binding DNA with a sequence-dependent energy. Sliding can thus be treated as 1D diffusion in an external-coordinate–dependent field. In earlier work [39, 40, 37], our group considered the sequence-dependent field as a random field with energies independently and normally distributed. The normal approximation is justified on the basis of closely matching the distribution of the protein-DNA binding energies computed using a popular position-weight matrix (PWM) approximation [41], which assumes that bound DNA base-pairs contribute independently and additively to the total binding energy, and that sufficiently many base pairs are present in a binding motif that, by the central limit theorem, the distribution of energies is normal. By averaging the mean-first-passage time for a 1D random walk over the normally distributed energies, one obtains

$$D_{1D} \propto \left(1 + \frac{\beta^2\sigma^2}{2}\right)^{1/2} e^{-\frac{7}{4}\beta^2\sigma^2} \tag{1.7}$$

where $\beta = (k_B T)^{-1}$ and σ^2 is the variance of the protein–DNA binding–energy distribution. The exponential factor falls off faster than the square-root factor grows for all σ, with

experimentally observed association rates consistent only with $\sigma \lesssim 1 - 2k_BT$. Consistent with this result most proteins with directly measured 1D diffusion coefficients have been shown to slide sufficiently fast with $\sigma \sim 1 - 2k_BT$ [42, 25, 22, 43, 44].

For proteins such as transcription factors that must not only locate and bind their target sites but also remain bound to effect their biological function (transcriptional activation in the case of TFs), stability of the protein–specific-DNA complex is an additional criterion that must be met in addition to rapid target localization. Our group has earlier demonstrated that the requirements of fast search and stability of the protein-DNA complex impose different and mutually exclusive constraints on σ (see Figure 1.2). The variance of the sequence-dependent binding energy σ determines not only the protein's diffusivity, but also the energy of the target site E_0, and hence the equilibrium occupancy of the target site P_{eq}:

$$P_{eq} = \frac{\exp(-E_0/k_BT)}{\sum_{i=1}^{M} \exp(-E_i/k_BT)}, \tag{1.8}$$

where energies E_i of individual sites are drawn from a normal distribution with variance σ^2 and the target site has the lowest energy in the genome

$$E_0 = \min_{i=1,\ldots,M} E_i \approx -\sigma\sqrt{2\log M}, \tag{1.9}$$

(with $M \approx 10^6$ bp for bacterial genomes). We can see from Equations 1.8 and 1.9 that $P_{eq} \gtrsim 0.25$ requires $\sigma \gtrsim 5k_BT$. From before, fast search, however, requires $\sigma \lesssim 1 - 2k_BT$. These two conditions' mutual unsatisfiability we term the *speed-stability paradox*.

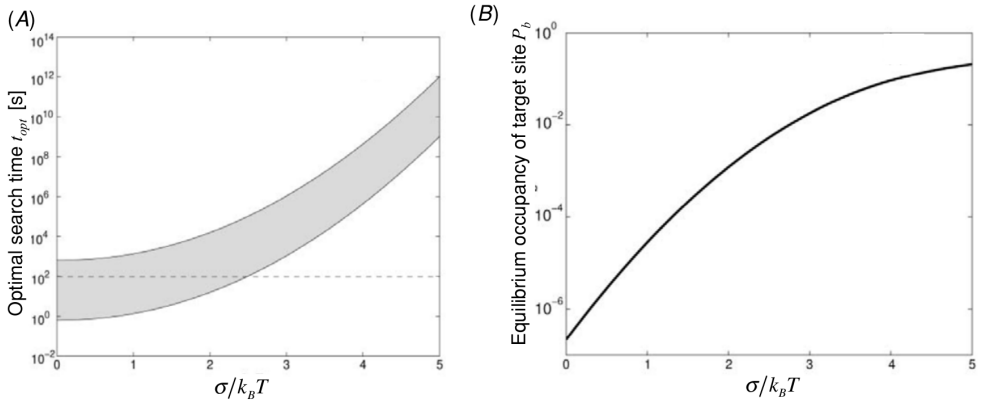


Figure 1.2: The speed-stability paradox. (**A**) The optimal search time for a single protein and a single target site on the entire bacterial genome. The lane corresponds to possible values for the search time depending on parameters of the model and assuming optimal 1D/3D partitioning. Fast searching is possible only if $\sigma < 1{\sim}2k_BT$. (**B**) The equilibrium occupancy of the target site that has the lowest possible energy among $M = 5 \times 10^6$ sites. High equilibrium occupancy (*i.e.* stability of the protein-DNA complex) requires $\sigma \gtrsim 5k_BT$. It is impossible to achieve both fast searching and stability if the classical model of sequence-dependent protein-DNA interactions applies.

1.3.4 A two-mode model

The model

To address the speed-stability paradox, we proposed model in which a protein may bind in two distinct (presumably conformational) modes [39, 37] (Figure 1.3) [1]. In the *search* or *sliding* mode, denoted as the **S** mode, the protein associates with the DNA such that $\sigma \lesssim 1 - 2k_BT$, and thus it can slide rapidly. In physical chemical terms, the **S**-mode protein and DNA associate through some combination of electrostatic and hydrophobic interactions as well as hydrogen bonding to the sugar-phosphate backbone. In the *recognition* or *reading* mode, denoted as the **R** mode, the protein's conformation is such that it more intimately interrogates the information-carrying parts of DNA, chiefly, the major groove. The complex

[1] For the sake of concision, we speak of states of the protein when really states of the protein-DNA complex are what is meant. Experimental studies have shown that when a DBP is on its target site, the DNA may be bent [45], or have a nucleobase extruded [46].

will be stabilized if the position and orientation of hydrogen-bonding moieties, π-interacting moieties, etc. in the protein and DNA are such that the interaction is favored, and will be destabilized by these moieties poorly complementing each other and as well as by steric clashes. Except for the infrequent positions where the protein is at its target site or a decoy, these moieties responsible for recognition that are brought into contact by the more intimate binding of the protein to DNA will not be well matched, and the free energy of the complex with the protein in the **R** mode will generally be greater than the free energy of the complex with the protein in the same position along the DNA but in the **S** mode. When the protein is in the **R** mode *and* is on the target site, however, it forms a very stable complex with DNA. Translocation in the **R** mode is considered negligible, either because of a large **R**-mode σ, or because of large energy barriers between positions on the DNA.

In the two-mode model, the total average search time is given by, instead of Equation 1.3,

$$\bar{t}_s = \frac{M}{\bar{n}} \frac{1}{P_f} (\tau_{1D}(1 + K_{R/S}) + \tau_{3D}). \tag{1.10}$$

where P_f is the probability of recognizing (not missing) the site upon sliding in its vicinity. The total search time thus requires a factor equal to the average number of times the global search needs to be repeated until recognition, *i.e.* $1/P_f$. It will be convenient to define τ_{1D} as the average time spent *sliding* in 1D, and since not all of the time the protein spends on DNA is spent in search mode, the τ_{1D} term needs to be multiplied by the factor $1 + K_{R/S}$, where $K_{R/S}$ is the equilibrium constant of the S$\rightleftarrows$R transition.

Structural evidence

Experimental evidence of distinct specific and non-specific binding modes of prokaryotic DBPs to cognate and non-cognate DNA, respectively, is found in NMR studies of *lac* repressor [45, 48], crystallographic measurements on the restriction enzymes BamHI [49] and

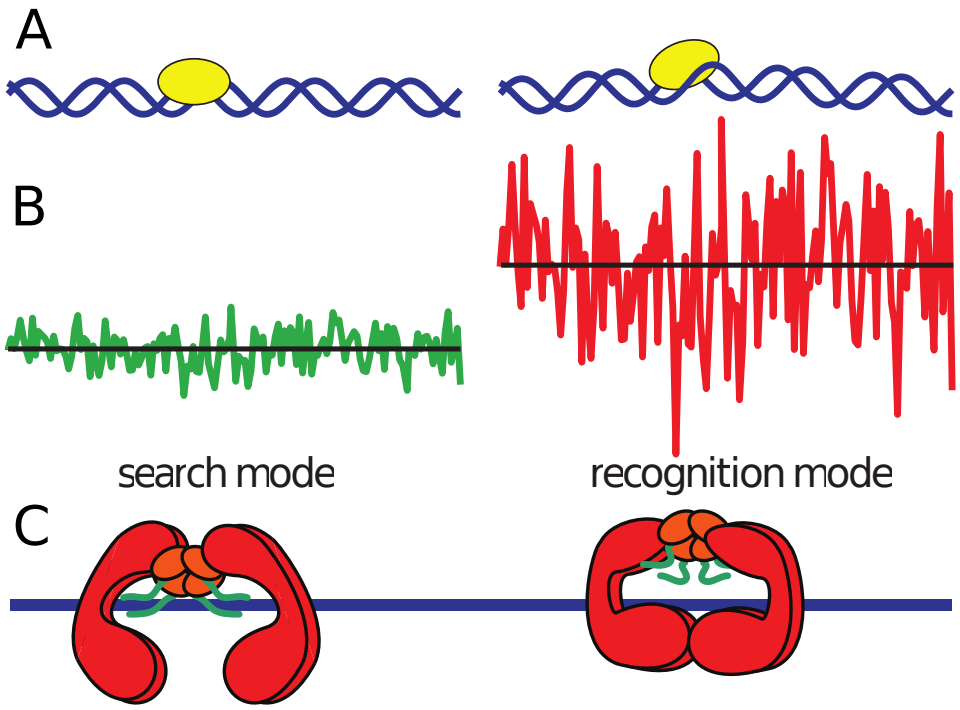


Figure 1.3: Energy landscapes and cartoons of proteins on DNA in search (**S**) and recognition (**R**) modes. **(A)**, **(B)**: In **S** mode, a generic protein (yellow) interacts chiefly with the DNA backbone and experiences a smooth landscape. In **R** mode, it interacts with the nucleobases, yielding a highly sequence-dependent landscape. **(C)**: Cartoon model for p53, based on EM data[47], indicates the domains responsible for the modalities: green C-terminal domain for the **S** mode; red core domain for the **R** mode. Tetramerization domain shown in orange. The color scheme matches that of Figure 3.4 in Chapter 3; N-terminal domain omitted.

BstYI [50], and scanning force microscopy on the *cro* repressor [51]. More recently, evidence of multiple binding modes in eukaryotic TFs have been found as well, and transcriptional activation by the yeast TF *Mbp1* has been shown to involve 1D sliding [52]. Multiple binding conformations have been identified from electron microscopy on p53 bound to an oligonucleotide containing a cognate sequence flanked by non-specific DNA [47]. Further support for a multi-mode model describing p53's interaction with DNA is provided by a single-molecule study of p53 truncation mutants that show that distinct domains—the C-terminal domain and the core domain—are responsible respectively for p53's sliding and recognition functionalities [53] (Figure 1.3C).

As will be discussed in the following chapter, the efficacy of the two-mode model of facilitated diffusion requires that transition between the two modes be sufficiently rapid. H-D exchange data on *lac* repressor suggest that conformational changes between the non-specific and specific conformations of the protein occur on the timescale of $10^{-5} - 10^{-3}$ sec [54, 55]. Earlier studies reported similar timescales and magnitudes of structural rearrangements in protein-DNA complex upon binding to a cognate site [56, 57] or for detection of damaged sites in DNA.

In the following chapter, I present the results of our investigation of the two-mode model. Particular attention is given to the implications of sequence-dependent binding energies in the **S** and **R** modes and whether the sequence-dependence between the two is correlated. The model gives predictions for the ranges of experimentally measurable parameters necessary to afford efficient target search.

Chapter 2

Two-mode model of protein-DNA interaction

2.1 Introduction

Experimental and theoretical developments leading up to the present work has been discussed in Chapter 1. Here, I present the results from two-mode models of increasing sophistication: a sequence-independent two-mode model (Figure 2.1A), a sequence-dependent two-mode model with uncorrelated search (**S**) and recognition (**R**) landscapes (Figure 2.1B), and a sequence-dependent two-mode model where disorder in the **S** mode is correlated with disorder in the **R** mode (Figure 2.8A). The correlation of the **S** and **R** landscapes gives a protein searching for its site an enhanced probability to transition into the **R** mode while on its target site, without the input of energy, and thus allows the protein to fold orders of magnitude slower than it otherwise would have to. This speed-up in search we call "kinetic pre-selection".

Simulations of random walks of a protein on DNA according to our model energy

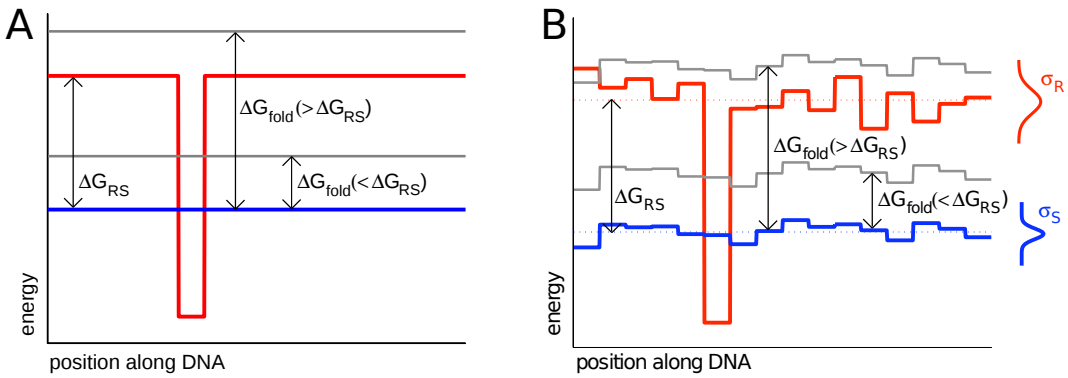


Figure 2.1: Energy landscapes in the vicinity of the protein's cognate site for (A) the sequence-independent two-mode model, and (B) the sequence-dependent two-mode model. In (A), the **S** landscape (blue) is flat, as is the **R** landscape (red) except at the cognate site. In (B), both landscapes are rugged, with standard deviation in their energies σ_S and σ_R respectively. ΔG_{RS} denotes the separation in energy between the two landscapes (between the respective means in (B)). The transition state (gray) between the **S** and **R** modes may be higher in energy than the **R** mode, or it may be less. In the latter case, the minimum folding barrier, ΔG_{fold}, is significant only at the cognate site and at occasional "traps", as for most positions x, the energy difference between the **R**- and **S** modes, $G_R(x) - G_S(x)$, exceeds ΔG_{fold}.

landscapes agree with analytical predictions. While the two-mode model in the absence of pre-selection allows for search that is nearly as fast as the ideal case, this requires that the protein fold faster than is observed experimentally for many proteins. Pre-selection allows both closer-to-ideal search efficiency as well as a more generous range of parameters compatible with efficient search; this range includes the folding rates that are too slow for the two-mode model without preselection. We further demonstrate that for given parameter values, there exists an optimum probability of folding into recognition mode upon visiting a site, that is, a point at which recognition is balanced between being sensitive and being discriminating. Our work solves the speed-stability paradox, and demonstrates the importance of conformational flexibility in protein-DNA interactions.

2.2 The sequence-independent two-mode model

2.2.1 Model

Although site-specific DNA-binding proteins are expected to exhibit sequence-dependence in their affinity to binding in the **R** mode, we find it instructive to consider a model in which both the **R** and **S** landscapes are completely flat, except for the **R** landscape at the target site, which has a deep well (Figure 2.1A). This model may in fact be largely accurate for the base-excision repair protein MutM [58], and possibly for DNA-damage-repair proteins generally that look for a rare and distinctive feature rather than for a DNA sequence, as do transcription factors. Regardless of the sequence-dependence or -independence of the **R** and **S** landscapes, applying to all versions of our two-mode model is the general equation for total average search time (Equation 1.10 reprinted for convenience):

$$\bar{t}_s = \frac{M}{\bar{n}} \frac{1}{P_f} (\tau_{1D}(1 + K_{R/S}) + \tau_{3D}). \tag{2.1}$$

In the sequence-independent model, $K_{R/S}$ is simply equal to e raised to the separation in energy between the **R** and **S** (flat) landscapes, $K_{R/S} = \exp(-\Delta G_{RS})$. ΔG_{RS} is positive and thus $K_{R/S} < 1$ when the **R** mode is less stable than the **S** mode. $K_{R/S}$ may also be considered equal to the ratio of the rate of folding from the **S** to the **R** mode, k_f, to the rate of the reverse transition, *i.e.* the unfolding rate k_u from the **R** to the **S** mode. The forward rate, k_f, is equal to barrier-less transition rate k_0 times e raised to the folding barrier ΔG_{fold}, and the reverse rate, k_u, is equal to k_0 times e raised to $-(\Delta G_{fold} - \Delta G_{RS})$ or to zero, whichever is greater (Figure 2.2). For simplicity, thermodynamic beta, $1/k_BT$, is omitted from equations and expressions–all energy parameters are in units of k_BT.

The other variable in the equation for the total average search time that introducing a second mode of binding adds to the one-mode model is the reciprocal of the probability to fold from **S** to **R** during a 1D search round that includes the target site, $1/P_f$. This,

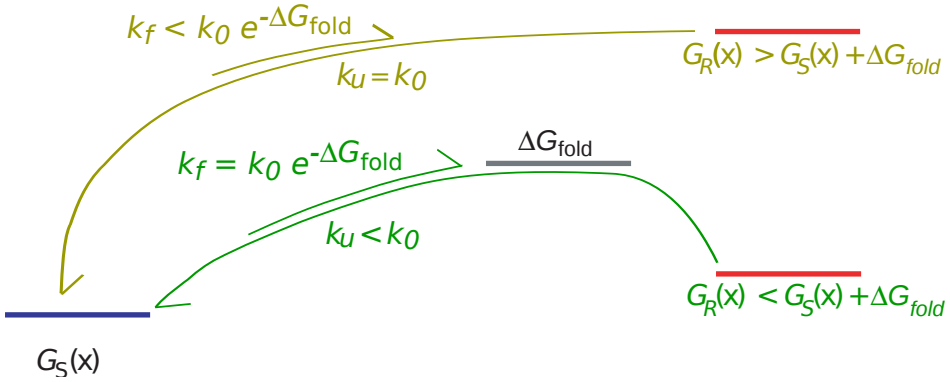


Figure 2.2: Relation of rates and in the two-mode model energies at a given position x on DNA. The folding rate, k_f, has a maximum of $k_0 \exp(-\Delta G_{fold})$, where k_0 is the rate of a zero-barrier transition, and ΔG_{fold} is the minimum folding barrier. This situation obtains when the energy in the **R** mode at position x, $G_R(x)$, is less than the energy in the **S** mode plus the minimum folding barrier (green arrows and labels). In this case, unfolding from the **R** back to the **S** mode has an activation barrier, and so the unfolding rate k_u is slower than k_0. If, on the other hand, the energy in the **R** mode at position x, $G_R(x)$, is greater than the energy in the **S** mode plus the minimum folding barrier (ochre arrows and labels), then the folding rate k_f is smaller than its maximum, while the unfolding rate is at its maximum.

the average number of times the protein must conduct a 1D search rounds over the target site, depends on the rate of **S**→**R** transition k_f and the average total residence time τ_{res} the protein spends on the target site while in the **S** state. If we consider the total residence time t_{res}, that is, sum of the durations of all the visits to a site during a 1D sliding round, as an exponentially distributed random variable, $1/P_f$ is simply

$$\frac{1}{P_f} \approx \frac{k_f + \tau_{\text{res}}^{-1}}{k_f} \tag{2.2}$$

As t_{res} is not exponentially distributed, however, accuracy requires a more complicated equation for $1/P_f$:

$$\frac{1}{P_f} = \frac{1}{\sqrt{\pi}} \frac{\exp\left(-\alpha^2\right)}{\alpha} \left(1 - Erf(\alpha)\right)^{-1} \ , \qquad \alpha = \frac{\tau_{res} k_f}{2} \tag{2.3}$$

The dimensionless parameter α is a measure of folding efficiency. See *Methods* subsection 2.5.3 for the derivation of Equation 2.3 and discussion of why Equation 2.2 is only an approximation. Note that τ_{res} is not the average time the protein stays on the site on each visit before it slides left or right, rather it is the average *total* amount of time of all such visits before the protein dissociates from this region of DNA. A simple estimation of the residence time is the time spent in the round of sliding divided by the number of sites visited. Since a protein makes $\sim n^2$ step while visiting $\sim n$ sites, the residence time can be approximated as

$$\tau_{\text{res}} \approx \frac{\tau_{1D}}{\bar{n}^2/\bar{n}} = \frac{\tau_{1D}}{\bar{n}}. \tag{2.4}$$

Equations (2.1), (2.3) and (2.4) allow us to calculate the average total search time. The no-sequence-dependence version of the two-mode model is fully described by two parameters beyond the one-mode model, ΔG_{RS} or k_u, and ΔG_{fold} or k_f.

2.2.2 Results

Since the two additional parameters that define the model are ΔG_{RS} and ΔG_{fold}, we focus on how these parameters affect the search time $\bar{t}_s$. Examining equation (2.3) shows that there will be a fast-folding regime where $\frac{k_f \tau_{\text{res}}}{2} \equiv \alpha \gg 1$, giving $1/P_f \approx 1$ and a search time independent of k_f, and a slow-folding regime where $\alpha \ll 1$, giving $1 < 1/P_f \propto 1/k_f$, equivalent to $\log(t_s) \propto \Delta G_{fold}$ (Figure 2.3A). The transition between regimes is found at the value of ΔG_{fold} that gives $k_f \approx \tau_{\text{res}}^{-1}/2$, which, using the same typical physical parameters for proteins we use in our simulations, is $= 4.0 k_B T$, corresponding to a minimum folding rate of $1.9 \times 10^5/s$ (derivation in *Methods* subsection 2.5.4).

The effect of ΔG_{RS} on t_s likewise exhibits two regimes. ΔG_{RS} affects $\bar{t}_s$ through the $\tau_{1D}(1 + K_{R/S})$ component of Equation (2.1). When $K_{R/S}(= \exp(-\Delta G_{RS})) \gg 1$, the protein spends much time unproductively in the **R** mode; this is the slow-sliding regime. As such, reducing $K_{R/S}$ allows significantly faster search times: $\bar{t}_s$ decreases exponentially with increasing ΔG_{RS} up to the point where $1 \approx K_{R/S}$ [1]. When $1 \gg K_{R/S}$, the protein spends neglible time in the **R** mode (except of course at the target site) and the system is in the fast-sliding regime. Placing the **R** mode being only a few $k_B T$ above the **S** mode suffices to keep the system in the regime favoring the **S** mode (Figure 2.3B). We will see that the point of transition between the two regimes is very different when disorder in the **S** and **R** modes is introduced.

If the model is parametrized with k_u rather than with ΔG_{RS}, we find that for a constant k_u there exists an optimum k_f (Figure 2.4). When k_f is larger than necessary for P_f to approach unity, it causes $K_{R/S}(= k_f/k_u)$ to grow, *i.e.* it causes the protein to waste time in the **R** mode more than is necessary to ensure that it recognizes its target site nearly

[1] provided that the duration of sliding rounds is substantially greater than the duration of 3D-diffusion rounds, which is the case for our simulations, and is observed experimentally [42, 53]

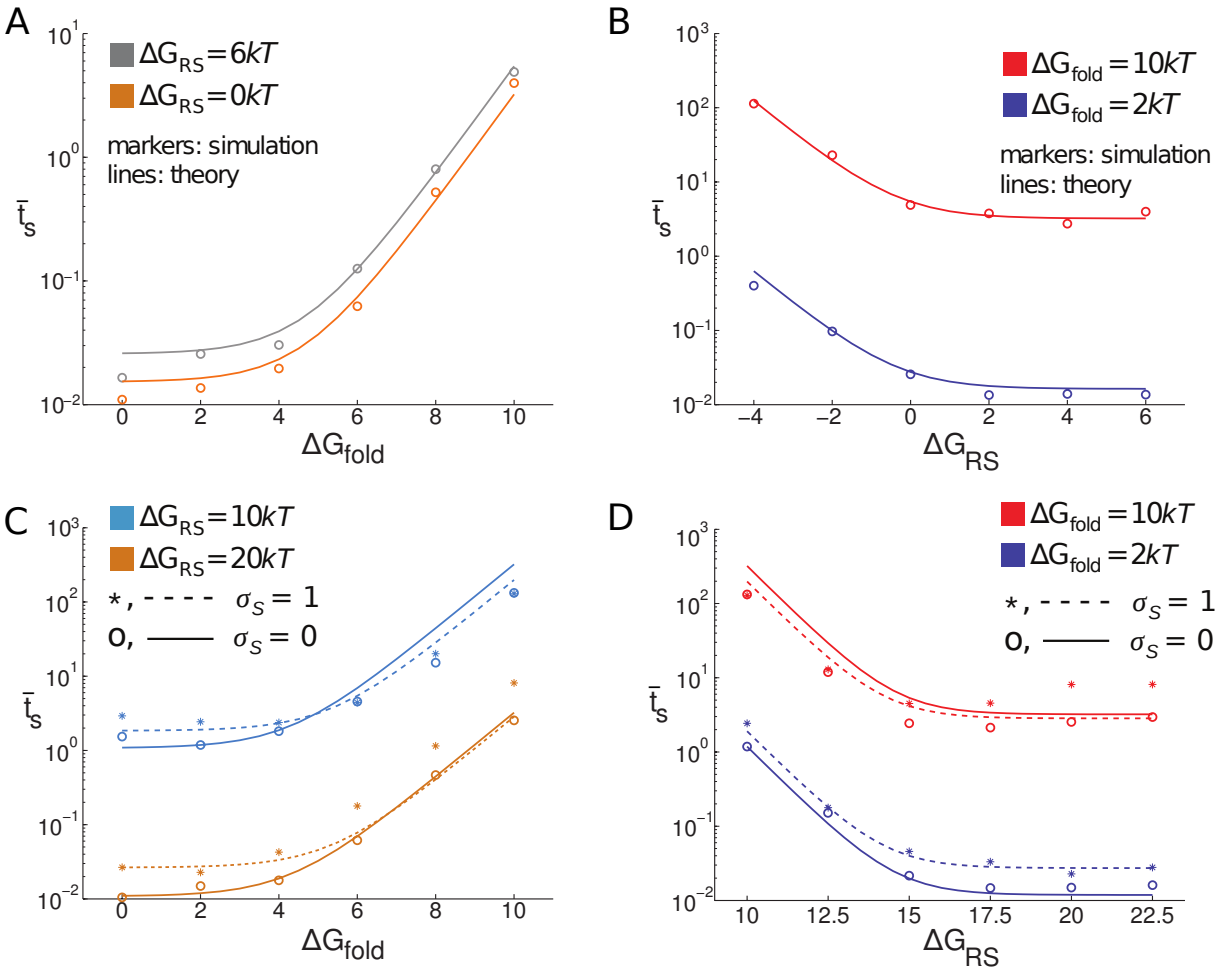


Figure 2.3: Two-mode model: Plots show search times $\bar{t}_s$, in seconds, as functions of the physical parameters ΔG_{fold} (plots 2A and 2C, on left) and ΔG_{RS} (plots 2B and 2D, on right), both in k_BT. Upper plots (2A and 2B) are for the model with no sequence-dependence disorder in energies and show only theoretical values; lower plots (2C and 2D) are for the model with disorder. The non-monotonicity of the simulated results in the sequence-dependent model is explained by translocation in the **R** mode immediately adjacent to the target site; we present a correction to the model that accounts for this behavior in subsection 2.3.2. In the interest of speed, simulations were run with a genome of length $M = 10^3$bp. As $\bar{t}_s \propto M$, an efficient search taking $\sim 10^{-2}$s in the simulations corresponds to a search time in a typical bacterial genome of10^6~10^7 bp of 10^1~10^2s.

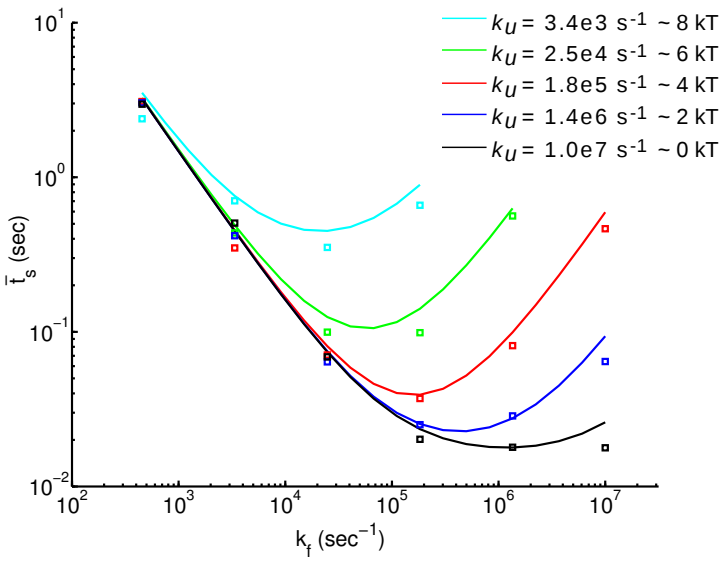


Figure 2.4: Mean search time as a function of k_f, with k_u held constant. Markers are simulated data; traces are analytical. Keeping a constant k_u while varying k_f is equivalent to keeping a constant $\Delta G_{fold} - \Delta G_{RS}$, while varying ΔG_{fold}. If folding is too fast, then the protein wastes time with unnecessary visits to the **R** mode.

every time it undergoes a sliding round that includes the site. When k_f is decreased, the factor representing delay owing to visiting the **R** mode, $1 + K_{R/S}$, asymptotically decreases to unity, but the factor representing delay owing failing to fold when encountering the target site, $1/Pf$, increases without limit (see Equation 2.3). The optimum is reached where:

$$k_f \approx \sqrt{\frac{k_u \bar{n}(\tau_{1D} + \tau_{3D})}{\tau_{1D}^2}} \tag{2.5}$$

(See *Methods* for derivation.)

The maximum unfolding rate k_u is the zero-barrier transition rate, which is equal to the rate of sliding one base–pair and thus $\approx (\tau_{1D}/n^2)^{-1}$. Substituting this value for k_u and Equation 2.5 and the resulting value of k_f into Equation 2.1 using the approximation of $1/P_f$ (Equation 2.2), and noting that $K_{R/S}$ must equal k_f/k_u, gives an optimal average

search time of

$$\bar{t}_{s,opt} = \frac{M}{\bar{n}}\tau_{1D}\left(r_\tau + \frac{1+r_\tau}{\sqrt{r_\tau \bar{n}}} + \frac{1}{r_\tau \bar{n}}\right), \quad r_\tau \equiv \frac{\tau_{1D}+\tau_{3D}}{\tau_{1D}} \tag{2.6}$$

The slowdown relative to an ideal one-mode model is loosely $\propto 1/\sqrt{\bar{n}}$. If the relative slowdown were a constant factor, it would imply that the protein needed to sample the **R** mode with some constant probability every time it visited a site, *i.e.* after every step on the DNA. If the relative slowdown were $\propto 1/\bar{n}$, that would imply that the protein needed merely to sample the **R** mode some average number of times per site among the $\bar{n}$ sites. Our intermediate result is reasonable considering that as $\bar{n}$ increases, the probability of sampling the **R** mode on a single visit may decrease and still offer a very good chance of sampling the **R** mode at least once. If the average absolute number (rather than the probability per visit) of excursions to the **R** mode, however, does not increase with increasing $\bar{n}$, however, a higher proportion of sites will be visited zero times in the **R** mode, and the protein risks failing to recognize its target site during the sliding round.

2.3 The sequence-dependent two-mode model, uncorrelated landscapes

2.3.1 Model

As described in Chapter 1, binding energies of protein-DNA complexes are in fact sequence-dependent. The free energy $G_R(x)$ at position $x \in 1, ..., M$ in the genome in the **R** mode is determined from energies calculated from a weight matrix, and approximates a Gaussian distribution with standard deviation $\sigma_R \gtrsim 5k_BT$. The mean of G_R is, as before, ΔG_{RS} above the average energy in the **S** mode. With respect to the **S** landscape, we now speak of an *average* energy because disorder may be introduced here as well, such that the

energies $G_S(x)$ are independent and normally distributed with standard deviation σ_S and are uncorrelated to the energies in the **R** mode.

The equation for the average search time, Equation (2.1), is still valid, although the expressions for some of its terms are more complicated. The diffusion coefficient for sliding, D_{1D}, was in the no-disorder model a free parameter, but now can be expressed as a function of σ_S (*Methods*, Equation 10 in [39]). The average time spent during a 1D diffusion round in the **S** mode, τ_{1D}, is also dependent on σ_S due to a lowering of energy in some positions and thus a shift in the 3D-1D equilibrium in favor of the **S** mode relative to the solution phase (*Methods*). The folding rate on the target site, k_f, is unchanged by adding disorder and remains $= k_0 \exp(-\Delta G_{fold})$. Because τ_{res} (Equation 2.4) depends on τ_{1D} and thus on σ_S, so also does $1/P_f$ (Equation 2.3). The expression for $K_{R/S}$ is derived from the random energy model (see *Methods*) and depends on ΔG_{RS}, σ_R, and σ_S:

$$K_{R/S} = \exp\left(\frac{\sigma_R^2 - \sigma_S^2}{2} - \Delta G_{RS}\right) \cdot \left(\frac{1 - \mathrm{Erf}\left(\frac{\sigma_R}{\sqrt{2}} - \sqrt{\log(M/\sqrt{2\pi}}\right)}{1 - \mathrm{Erf}\left(\frac{\sigma_S}{\sqrt{2}} - \sqrt{\log(M/\sqrt{2\pi}}\right)}\right) \tag{2.7}$$

In the limit of large M, equation 2.7 reduces to $K_{R/S} = \exp(\frac{\sigma_R^2 - \sigma_S^2}{2} - \Delta G_{RS})$. The non-exponential factor accounts for lack of sampling in the low-energy tails of the distributions of energies—as $M \to \infty$, the tails are increasingly well sampled and the factor approaches unity for all finite values of σ_R and σ_S.

2.3.2 Results

Fast- and slow-sliding regimes

The most significant effect of the disorder is to change the value of ΔG_{RS} that marks the transition between the fast-sliding and the slow-sliding regimes (compare Figures 2.3B and 2.3D). The point of transition between the regimes in the no-disorder model lies at $\Delta G_{RS} = 0 k_B T$, while for the full model it lies at $\sim 15 k_B T$ for both theoretical calculations

and our simulated data (see *Methods* subsection 2.5.5 for derivation). This effect is explained by noting that $K_{R/S}$ has an $\exp(\sigma_R^2/2)$ dependence as well as an $\exp(-\Delta G_{RS})$ dependence, which means that ΔG_{RS} must be much greater to achieve $K_{R/S} \ll 1$. Assuming the random energy model for the energies of the **R** mode, a genome of more than a few hundred base pairs will have "traps" that are as easy to fold on as is the target site and have a large barrier to unfold on. The time spent in these traps can be decreased by raising their energy in the **R** mode, that is, by increasing ΔG_{RS}. This can be achieved by (i) making the protein somewhat unstable and hence favoring the partially unfolded **S** mode, and/or (ii) making the **R** conformation stressed, *e.g.* by deformation of the DNA. These phenomena are observed in the partially unfolded conformation of Lac repressor on a non-cognate DNA sequence obtained by NMR [45], with a sharp DNA bend in the cognate complexes [59], and with the high heat capacity of the conformational transition reported for several DNA-binding proteins [57]. In addition to panels (C) and (D) in Figure 2.3, the data may be visualized as contour plots in Figure 2.5.

The **R** mode cannot be destabilized beyond a certain point, however, or the energy of the protein while folded on its target site will be too great for it to be stably bound there. ΔG_{RS} must be low enough that the protein remains bound sufficiently long to perform its biological function. Thus, the stability criterion imposes an upper limit on ΔG_{RS} while the speed criterion imposes a lower limit. How stable the cognate-site protein-DNA complex must be depends of course on the particular system; for illustrative purposes, we show in the contour plots (Figure 2.5) dotted lines corresponding to the maximum ΔG_{RS} such that separation in energy between the target site in the **R** mode and the target site in the **S** mode is greater than the separation between the **S** mode and solution (*i.e.* τ_{1D}).

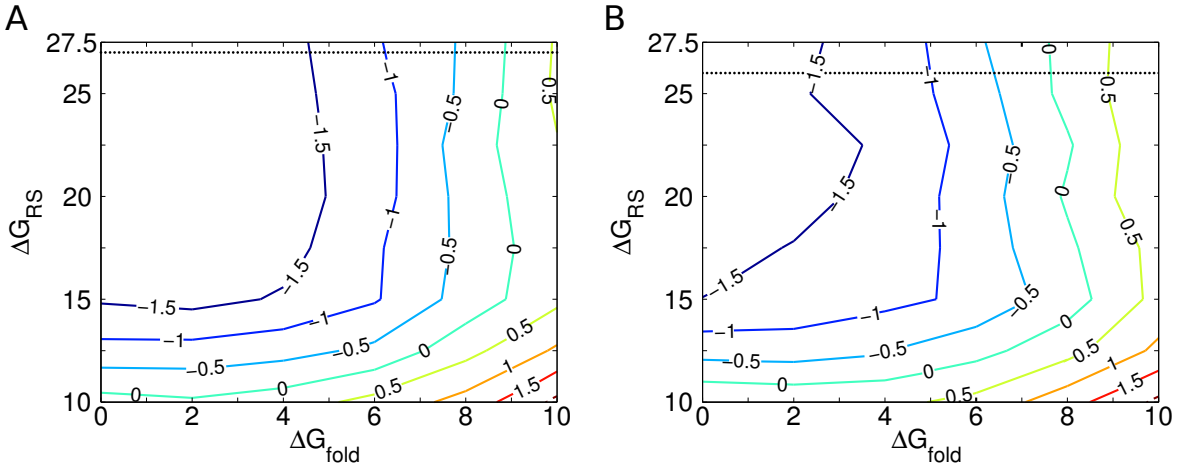


Figure 2.5: Contour plots of the mean search times as a function of ΔG_{fold} and ΔG_{RS}. Contours are labeled according to the base-10 logarithms of their average search times in *sec*. Data is for sequence-dependent uncorrelated landscapes. **A**: $\sigma = 0k_BT$. **B**: $\sigma = 1k_BT$. The dotted black lines denote the values of ΔG_{RS} such that the separation in energy between **R**-mode complex on the target site and the **S**-mode complex on the site is equal to the average separation between the **S**-mode complex and solution. Panels **A** and **B** differ the most in the upper-left quadrant (fast-sliding, fast-folding regime) and are scarcely distinguishable in the lower-right quadrant (slow-sliding, slow-folding regime). The regime-dependent effects of introducing $1k_BT$ of disorder into the **S** landscape are summarized in Figure 2.6.

Dependence of search time on σ_S

We examined an **S** landscape that was entirely flat as well as one where the energy of each position was drawn from a normal distribution with $\sigma_S = 1k_BT$. The introduction of ruggedness into the **S** landscape, unsurprisingly, can increase the average search time. The effect of σ_S on $\bar{t}_s$ depends, however, on the regime in ΔG_{RS}–ΔG_{fold} phase space (Figure 2.6).

The factors in the master equation for $\bar{t}_s$ (Equation 2.1) that contribute to a σ_S

	fast-folding ($k_f \gg \tau_{res}^{-1}$) ($\Delta G_{fold} < 4.0$ kT)	slow-folding ($kf \ll \tau_{res}^{-1}$) ($\Delta G_{fold} > 4.0$ kT)
fast-sliding ($K_{R/S} \ll 1$) ($\Delta G_{RS} > 15$ kT)	$\exp(9/8\sigma^2)$	$\exp(5/8\sigma^2)$
slow-sliding ($K_{R/S} \gg 1$) ($\Delta G_{RS} < 15$ kT)	$\exp(1/2\sigma^2)$	$\exp(0)$

Figure 2.6: Regime-dependence of σ_S. Entries are the factor by which $\bar{t}_s$ is increased as a function of σ_S in the various regimes. The regime-dependent effects of σ_S are seen in Figure 2.5, where the greatest difference between panels **A** and **B** is found in the upper-left quadrants (fast-folding, fast-sliding regime), while the differences between panels in their lower-right corners (slow-folding, slow-sliding regime) are barely discernable.

dependence irrespective of the regime are $\bar{n}$ in the denominator and τ_{1D} in the numerator. Since the protein spends more time in lower-energy positions on DNA, τ_{1D} is proportional to $\exp(\sigma_S^2/2)$ (see *Methods*). The σ_S-dependence in $\bar{n}$ owes both to dependence in D_{1D} (Equation 1.7) and in τ_{1D}, as $\bar{n} = \sqrt{4D_{1D}\tau_{1D}}$ (Equation 1.4), and is proportional to $\exp(-\frac{5}{8}\sigma_S^2)$. Combined, these factors give $\bar{t}_s \propto \exp(\frac{9}{8}\sigma_S^2)$.

In the slow-sliding regime, that is, when ΔG_{RS} is not large enough to prevent wasteful visits to the **R** mode, our theory predicts that since $K_{R/S} \gg 1$, Equation 2.7 should contribute a $\exp(-\sigma_S^2/2)$ factor to $\bar{t}_s$. When $K_{R/S} \ll 1$, however, unnecessary visits to the **R** mode are rare and this factor vanishes.

In the slow-folding regime, where $k_f \ll \tau_{res}^{-1}/2$, *i.e.* $\alpha \ll 1$, the factor $1/P_f$, representing the average number of 1D sliding rounds the protein must undergo in the vicinity of the target site before recognizing it, reduces to $\frac{2/\sqrt{\pi}}{\tau_{\mathrm{res}} k_f}$. Using Equations 2.4, 1.4, and 1.7, it can be seen that this factor is proportional to $\exp(-\frac{5}{8}\sigma_S^2)$. In the fast-folding regime, however, where $1/P_f \approx 1$, it contributes no such σ_S-dependence.

When the system is in the slow-sliding and slow-folding regime, the σ_S dependences of all the factors cancel, and $\bar{t}_s$ is independent of σ_S. This can be seen on the right side of Figure 2.3C, in the upper traces and markers, corresponding to $\Delta G_{R/S} = 10 k_B T$, which is a smaller separation than is necessary for $K_{R/S} \ll 1$ and thus within the slow-sliding regime. Where ΔG_{fold} is large, the markers and traces corresponding to $\sigma_S = 1 k_B T$ and those corresponding to $\sigma_S = 0 k_B T$ converge. On the left side of Figure 2.3C, however, ΔG_{fold} is small and folding on the target site is fast. The regime-independent dependence of $\bar{t}_s$ on σ_S, $\bar{t}_s \propto \exp(\frac{9}{8}\sigma_S^2)$, is only partially canceled, resulting in an $\exp(\frac{1}{2}\sigma_S^2)$, dependence. Our theory thus predicts that in the fast-folding, slow-sliding regime, increasing σ_S from 0 to 1 $k_B T$ should result in an $e^{1/2} = 1.6$-fold increase in average search time, in agreement with simulations (blue markers in Figure 2.3C).

In the fast-sliding regime, $K_{R/S} \ll 1$, *i.e.* the protein wastes little time in the **R** mode. If folding is also fast, then the only non-vanishing σ_S-dependent factors in the equation for $\bar{t}_s$ (Equation 2.1) are the regime-independent ones, and increasing σ_S/k_BT from 0 to 1 should result in an $e^{9/8}$-fold, or half an order of magnitude, increase in average search time. This is indeed observed on the left side of Figure 2.3C (orange markers). If folding is not fast, however, the $\exp(\frac{9}{8}\sigma_S^2)$ dependence is partially canceled by the $1/P_f$ factor, and increasing σ_S/k_BT from 0 to 1 should result in an $e^{5/8} = 1.9$-fold increase in average search time, which is borne out by simulations (orange markers in Figure 2.3C, right side (ΔG_{fold} is large)).

Sliding in the R mode

If the protein can translocate in the **R** mode, then it can reach its target site by sliding into it already folded in addition to sliding to it in the **S** mode and then folding. We thus introduce k_f^{eff}, the effective folding rate, which is the sum of the folding rate on the target site and at the sites immediately to the left and right of it[2].

The barrier to fold at the target site ($x = 0$) $= \Delta G_{fold}$. At sites immediate adjacent to the left and right ($x =$ -1 and +1), the barrier to fold may be higher if $G_R(x) - G_S(x) > \Delta G_{fold}$. Since $G_S(x) = \langle G_R \rangle - \Delta G_{RS}$, we have:

$$\begin{aligned} k_f^{\text{eff}} &= k_f + \\ & \quad k_0 \exp\left[-\max(\Delta G_{fold}, G_R(-1) - \langle G_R \rangle + \Delta G_{RS})\right] + \\ & \quad k_0 \exp\left[-\max(\Delta G_{fold}, G_R(+1) - \langle G_R \rangle + \Delta G_{RS})\right] \end{aligned} \tag{2.8}$$

Replacing k_f with k_f^{eff} in the expression for the total search time (Equations 2.3 and 2.1)

[2]It is also possible, of course, that the protein could have slid in the **R** mode from farther away, but this would only contribute significantly to k_f^{eff} if there were a funnel around the target site, which our model does not assume

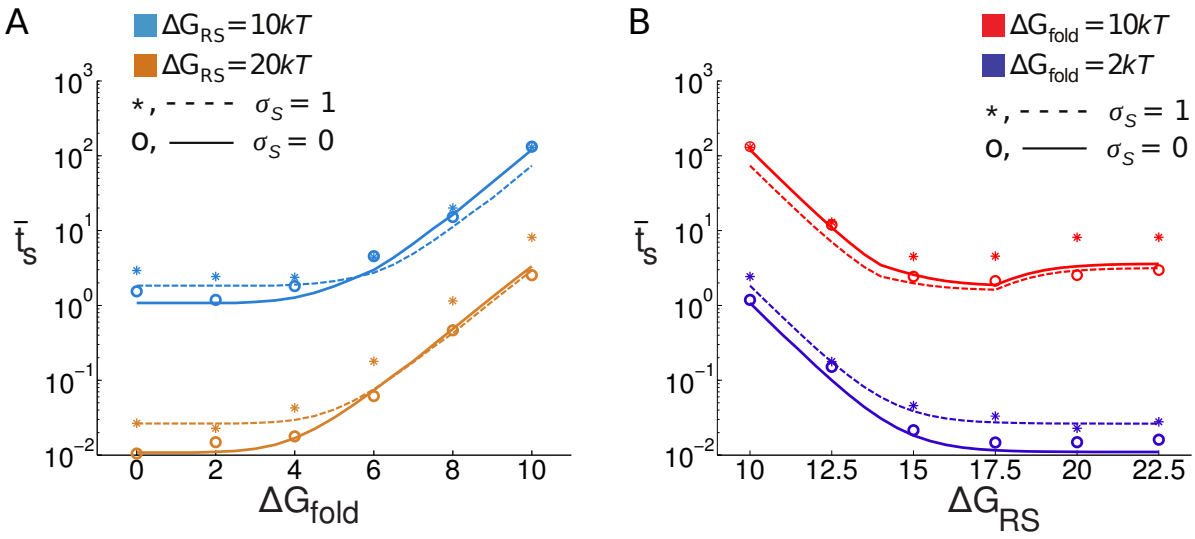


Figure 2.7: Two-mode model, sequence-dependent landscapes: Traces show $\bar{t}_s$ in seconds as a function of (A) ΔG_{fold}, and (B) ΔG_{RS}, both in k_BT, with analytical results (traces) from theory modified according to Section 2.3.2, "Sliding in the **R** mode". Markers are simulations and are identical to those in Figure 2.3C,D.

brings our theoretical predictions for the total average search time better in line with simulated results (Figure 2.7). Most importantly, it accounts for the non-monotonicity observed in the markers for $\Delta G_{fold} = 10k_BT$ in Figure 2.7B. As discussed in Section 2.2, the average total search time is sensitive to the folding rate k_f only when in the slow-folding regime, which is found at large ΔG_{fold}. Thus while in Equation 2.8, k_f^{eff} differs from k_f the most when ΔG_{fold} is small, this difference will not be reflected in the total search time, and the k_f^{eff} correction will only be important at small ΔG_{fold}. This can be seen by examining the theoretical traces in Figure 2.7B and Figure 2.3D and seeing that for $\Delta G_{fold} = 2k_BT$ they are nearly indistinguishable, while for $\Delta G_{fold} = 10k_BT$ they are qualitatively different. At large ΔG_{fold}, the total average search time is shortened for small ΔG_{RS}, which is expected from Equation 2.8. The mechanism of folding next to the target site and then stepping into it is effective only when $G_R(\pm 1)$ is not too high relative to $G_S(\pm 1)$, and so at larger values of ΔG_{RS} does not contribute significantly to the speeding-up of the search process.

Comparison with experiment

As discussed earlier, when $P_f \approx 1$, search is efficient as the protein rarely fails to fold during 1D-sliding rounds in which it reaches its target. This requires $\tau_{\rm res}k_f/2$ ($\equiv \alpha$) ≥ 1 (Equation 2.3). Experimental measurements of the rates of conformational transitions, k_f give a range of values: for *lac* repressor, k_f is estimated to be 10^3~10^5 $\rm s^{-1}$) [55, 54], and for the extrusion of base-pairs by lesion-binding proteins, $k_{\rm open}$~$10 - 100$ $\rm s^{-1}$) [60]. Depending on the system, then, $\tau_{\rm res}$ must be at least 10^5 s, in many instances it likely must be even greater.

Absent pre-selection, $\tau_{\rm res}$ can be estimated by Equation 2.4; it can be determined from any two of τ_{1D}, $\bar{n}$, and D_{1D}. *In vitro* single-molecule experiments that have estimated these values give diffusion coefficients $\approx$ 1–5 $\times 10^6$ $\rm bp^2/\mathit{s}$ [61, 22, 43] within an order of magnitude of the theoretical limit of $\approx 10^7$ [62]. Observed values of $\bar{n} \approx$ 50–200 bp[61, 22, 63, 21], and so we obtain a residence time $\tau_{\rm res} \approx$ 1–3 $\times 10^{-6}$ sec for the theoretical limiting D_{1D}, or $\approx 10^{-5}$–10^{-6} sec for experimental D_{1D}. Efficient search with these values of $\tau_{\rm res}$ require, in the former case, a folding rate k_f on the order of 10^6 $\rm s^{-1}$, and at the longer-$\tau_{\rm res}$ end of the latter case to $k_f \approx 10^5 \rm s^{-1}$, in agreement with our simulations and theory. These folding rates are at the upper limit of experimental measurements [55, 54], and suggest that all but the fastest-transitioning protein-DNA complexes will not be able to achieve efficient search on uncorrelated **S** and **R** landscapes.

In vivo experiments from which the researchers were able to infer the parameters that are necessary to determine $\tau_{\rm res}$, and thus the minimum k_f for efficient folding, are fewer, and tend to have somewhat higher implicit estimates of $\tau_{\rm res}$. Elf *et al.* performed single-molecule measurements on *lac* repressor in *E. coli*, and estimated 0.3ms$< \tau_{1D} <$ 5ms and $D_{1D} = 4 \times 10^5$ $\rm bp^2/s$, which gives $\tau_{\rm res} = 1$~6×10^{-5}s [42]. Very recently, Larson *et al.* imaged GFP fusions of the native transcription factor Mbp1 in yeast [52] and found a larger

τ_{1D} of 0.8s, and a diffusion coefficient of $D_{1D} = 5 \times 10^5$ bp^2/s, resulting in $\tau_{\text{res}} = 6 \times 10^{-4}$s. This value of τ_{res} is thus far the most permissive of slower folding.

2.4 The sequence-dependent two-mode model, correlated landscapes

2.4.1 Model

The models discussed thus far, as well as all but one system experimentally studied, require that proteins be able to fold faster than ~10^5s^{-1} for search to be efficient. In other words, k_f must be greater than this for $1/P_f$ to approach unity. This value for k_f, however, is the upper limit for experimentally observed folding rates of site-specific DBPs.

A naive attempt to counteract a small k_f would be to increase τ_{res}, the average total time per sliding round spent on given site (see Equation 2.3). Recalling from Section 2.2 that $\tau_{\text{res}} \approx \tau_{1D}/\bar{n}$, and substituting Equation 1.4 into Equation 2.4, we have

$$\tau_{\text{res}} \approx \sqrt{\frac{\tau_{1D}}{4D_{1D}}} \tag{2.9}$$

Greater residence times allow the protein more time to fold when on the target site, but increasing τ_{res} by a given factor comes at the price either of slowing the protein's random walk (i.e., decreasing D_{1D}) or of lengthening the duration of a sliding round, τ_{1D}, by that factor squared.

We have identified, however, a mechanism by which τ_{res} and thus P_f may be increased without compromising the speed of the protein's search. If disorder in the **S** mode is correlated with disorder in the **R** mode (Figure 2.8A), then the large energy minimum at the target site in the **R** mode corresponds to a small energy minimum at the target site,

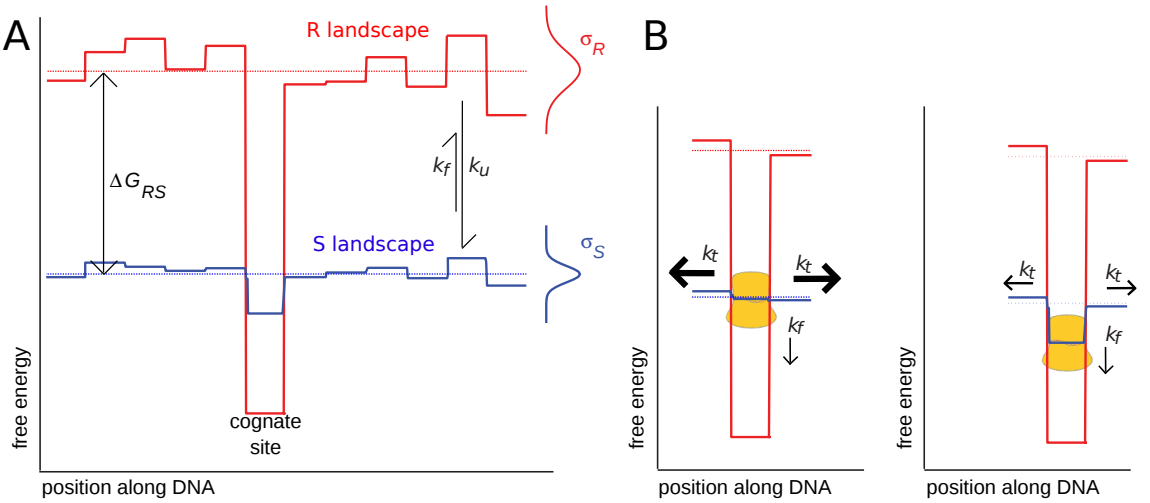


Figure 2.8: The two-mode model with kinetic preselection. (A) The **R** and the **S** landscapes have the same degree of disorder as in the two-mode model without preselction (Figure 2.1B), but the disorder is correlated. This causes a small potential well in the **S** landscape on the target site (B, right), in constrast to the absence of such a well in the uncorrelated model (B, left). The well reduces the translocation rates k_t away from the target site, making folding on the target site relatively more favored.

$(x = 0)$, in the **S** mode (Figure 2.8B). This energy minimum increases τ_{res} such that

$$\tau_{\text{res, pre-selection}} = \tau_{\text{res}} \exp\left(\frac{\sigma_S}{\sigma_R}(\langle G_R \rangle - G_R(0))\right) \tag{2.10}$$

This mechanism, which we term *kinetic pre-selction*, increases τ_{res} locally on its target site, allowing it more time to undergo conformational transitions preferentially on sites where such transitions are productive. Thus, the range of values of k_f necessary to achieve $1/P_f \approx 1$ may be expanded downward. Notably, kinetic pre-selection is not a proofreading mechanism, which would require some energy consumption to improve specificity of recognition. Rather, pre-selection increases both on- and off- rates of binding to the target site, thus having no effect on equilibrium binding and requiring no energy consumption.

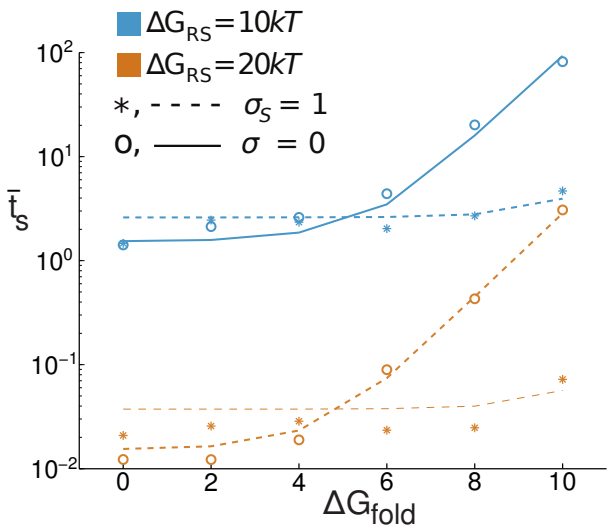


Figure 2.9: Two-mode model, correlated sequence-dependent landscapes: Mean search times $\bar{t}_s$ in seconds as functions of ΔG_{fold} in k_BT, with **S** and **R** landscapes correlated. Introducing 1 k_BT of correlated disorder into the **S** landscape allows for efficient search at much higher folding barriers. In the uncorrelated case (Figure 2.3C), the transition between fast-folding regimes (low ΔG_{fold}, left side of plots) and slow-folding regimes (high ΔG_{fold}, right side of plots) lies at $4.0k_BT$, equivalent to $k_f = 1.9 \times 10^5 s{-}1$. In the correlated case, folding barriers as large as $8.9k_BT$, equivalent to $k_f = 1.3 \times 10^3 s{-}1$. Aside from the correlation of the landscapes, all physical parameters are the same as in the uncorrelated case.

2.4.2 Results

The most striking consequence of correlating disorder in the **R** and **S** modes is that the boundary between the slow-folding and fast-folding regimes is shifted such that the rate of folding into the **R** mode, k_f, can be two orders of magnitude slower and still allow for efficient search and recognition. Comparing Figures 2.3C and 2.9, one sees that the maximum ΔG_{fold} for efficient folding increases from $4.0k_BT$ to $8.9k_BT$, corresponding to a decrease in the minimum folding rate from $1.9 \times 10^5/s$ to $1.3 \times 10^3/s$. Importantly, this two-order-of-magnitude decrease in the minimum folding rate for efficient search allows experimentally observed folding rates for the *lac* repressor [55, 54] to fall within the fast-folding regime.

We considered in our simulations only **S** landscapes with $\sigma_S = 1k_BT$ and, for reference and comparison, $0k_BT$. While increasing σ_S typically leads to greater mean search times, as discussed in Section 2.3.2, for proteins that, owing to steric or other reasons, cannot achieve a k_f rapid enough to fall within the fast-sliding regime with $\sigma = 1k_BT$, σ_S greater than $1k_BT$ may be optimal. Analytically, a σ_S of $1.5k_BT$, for example, would decrease the minimum k_f for efficient folding another order of magnitude, to $\sim 100/s$, at the expense of a reduced 1D diffusion coefficient and more-time-consuming 1D search rounds.

2.5 Methods

2.5.1 Simulations

To test the two-mode model, I implemented Monte Carlo simulations over a range of parameters, with and without pre-selection. I used the Gillespie algorithm [64], in which the rates, probabilities, and times of moves are given by:

$$\text{rate of move } m = k_m = \tau_0^{-1} \exp\big(-\Delta G^{\ddagger}(m)\big) \tag{2.11}$$

$$\text{prob}(m) = \frac{k_m}{\sum_i k_i} \qquad \Delta t = \frac{p}{\sum_i k_i} \tag{2.12}$$

where $\Delta G^{\ddagger}(m)$ is the energy barrier for move m, τ_0 is the mean attempt time or the average time to make move over a zero energy barrier, and $p \sim \text{Exp}(1)$, the standard exponential distribution. The sums are over all possible moves for the protein in its current state and position.

Each run of the simulation ran until the protein found its target site. For the sake of speed, a genome length M of 10^3 rather than $10^{6\sim9}$ was used. The physical parameters G_{3D}, ΔG_{RS}, σ_R, σ_S, and k_f, as well as whether the **S** and **R** landscapes were correlated (whether kinetic pre-selection was "on"), determine all the relevant energies. The energies

in turn govern the rates and probabilities of various moves, of which four kinds are possible:

Translocate :	Protein steps to the right (+) or left (−)	
	$\Delta G^{\ddagger} = \max\left(G(x \pm 1) - G(x)\ , 0\right)$	
Fold :	$\Delta G^{\ddagger} = \Delta G_{fold}$	if $G_R(x) < G_S(x) + \Delta G_{fold}$
	$\Delta G^{\ddagger} = G_R(x) - G_S(x) + \Delta G_{fold}$	if $G_R(x) > G_S(x) + \Delta G_{fold}$
Unfold :	$\Delta G^{\ddagger} = G_S(x) - G_R(x) + \Delta G_{fold}$	if $G_R(x) < G_S(x) + \Delta G_{fold}$
	$\Delta G^{\ddagger} = 0$	if$G_R(x) > G_S(x) + \Delta G_{fold}$
Dissociate :	$\Delta G^{\ddagger} = G_{3D} - G_S(x)$	Not allowed from the R state

ΔG_{fold} is the energy equivalent of k_f. It is the folding barrier on the target site and the minimum fold barrier generally; it equals $-\log(k_f \tau_0)$. $G_R(x)$ is the energy at position x of the **R** landscape; the energies are normally distributed with a mean of ΔG_{RS} above the **S** landscape, and standard deviation σ_R. Similarly, $G_S(x)$ is the energy at position x of the **S** landscape, with mean ΔG_{RS} below the **R** landscape, and standard deviation σ_S. G_{3D} is the free energy of the protein in solution, and is set such that $G_{3D} - \langle G_S \rangle = 10 k_B T$.

In the interest of speed, a genome length of 1000 was used. This time-saving device required a correction to $K_{R/S}$ (*Methods* 2.5.5). $G_R(x)$ was obtained from scoring the vicinity of a binding site for the *E. coli* transcription factor *purR* in the *E. coli* genome using a position-weight-matrix (PWM) approximation for binding energy.

2.5.2 D_{1D} and τ_{1D}

The 1D diffusion coefficient is given by[39]

$$D_{1D} \simeq \frac{1}{2\tau_0}(1+\sigma^2/2)^{1/2}\exp(-\frac{7}{4}\sigma^2) \tag{2.13}$$

where τ_0 is the effective attempt period; its corresponding rate is k_0. A value of $\tau_0 = 10^{-7}s$ is chosen, which, when $\sigma = 0k_BT$, yields a diffusion coefficient equal to the maximum theoretical value for a typical globular protein tracking the DNA helix as it slides [62], discussed further in Part II.

The average lifetime of a protein on DNA in the **S** mode equals τ_{1D}, which is given by

$$\tau_{1D} = \tau_0 \exp(G_{3D} - \langle G_S \rangle + \sigma_S^2/2) \tag{2.14}$$

where G_{3D} is the average free energy of binding non-specifically to DNA, *i.e.* $G_{3D} - \langle G_S \rangle$ is the average difference in energy between the protein being in solution and being bound non-specifically in the **S** mode.

2.5.3 Derivation of $1/P_f$

The probability that one event occurs before a second event is

$$P_1 = \int_0^{+\infty} f_1(t)[1 - F_2(t)]dt \tag{2.15}$$

where f_1 is the probability density function (PDF) of the waiting time for the first event and F_2 is the cumulative distribution function (CDF) of the second event. In our system, folding is a simple exponential process, so $f_1 = k_f \exp(-k_f t)$. The sum of the waiting times to translocate away from a site, however, is not exponentially distributed. The residency time on a site (of which τ_{res} is the average) is proportional the square root of the residency time on the DNA between jumps, which *is* exponentially distributed. Combining Equations

1.4 and 2.4 and writing them in terms of times instead of time constants gives us

$$t_{res} = \sqrt{t_{1D}/4D_{1D}} \tag{2.16}$$

Substituting t_{res} into a PDF for t_{1D} and integrating with respect to t_{res} from 0 to t gives us a CDF for t_{res} of $1 - \exp(-\frac{4D_{1d}}{\tau_{1D}}t^2)$, so Equation 2.15 becomes

$$P_f = \int_0^{+\infty} k_f \exp(-k_f t) \left[\exp\left(-\frac{4D_{1d}}{\tau_{1D}}t^2\right)\right] dt \tag{2.17}$$

Writing $4D/\tau_{1D} = \rho$ and changing the variable of integration to $y = \sqrt{\rho}t + \frac{k_f}{2\sqrt{rho}}$ gives us

$$P_f = \frac{k_f}{\sqrt{\rho}} \exp\left(\frac{k_f^2}{4\rho}\right) \int_0^{+\infty} \exp(-y^2) dy \tag{2.18}$$

whose reciprocal, $1/P_f$, evaluates to

$$\frac{1}{P_f} = \frac{2}{\sqrt{\pi}} \frac{\exp -\tau_{res}^2 k_f^2/4}{\tau_{res} k_f} \left(1 - Erf\left(\frac{\tau_{res} k_f}{2}\right)\right)^{-1}, \qquad \alpha = \frac{\tau_{res} k_f}{2} \tag{2.19}$$

When $\alpha \gg 1$ (fast-folding regime) this reduces to $1/P_f = 1$, and when $\alpha \ll 1$ (slow-folding regime) this reduces to $1/P_f = \frac{2/\sqrt{\pi}}{\tau_{res} k_f} = \sqrt{\pi}/\alpha$.

2.5.4 Points of transition between slow-folding and fast-folding regime

The value of ΔG_{fold} that marks the transition between slow-folding and fast-folding regimes is determined by setting the dimensionless parameter $\alpha = 1$ (*Methods*, subsection 2.5.3), and solving for ΔG_{fold}. The following applies equally to the no-disorder model and to the sequence-dependent model with $\sigma_S = 0$.

$$\alpha = \frac{k_f \tau_{\text{res}}}{2} = 1 \tag{2.20}$$

Recalling that $k_f^{-1} \equiv \tau_0 \exp(\Delta G_{fold})$, and substituting into this relation Equations 1.4, 2.14, and 2.13, we have

$$\begin{aligned} \tau_0 \exp(\Delta G_{fold}) &= \tau_{\text{res}}/2 \\ &= \sqrt{\frac{\tau_{1D}}{16 D_{1D}}} \\ &= \sqrt{\frac{\tau_0^2}{8}(1+\frac{\sigma_S^2}{2})^{1/2} \exp((G_{3D} - \langle G_S \rangle) + \frac{9}{4}\sigma_S^2)} \end{aligned} \tag{2.21}$$

With $\sigma_S = 0$, this reduces to:

$$\begin{aligned} \tau_0 \exp(\Delta G_{fold}) &= \sqrt{\frac{\tau_0^2}{8} \exp(G_{3D} - \langle G_S \rangle)} \\ \Delta G_{fold} &= \frac{1}{2}\left(\log(\frac{1}{8}) + (G_{3D} - \langle G_S \rangle) \right) \\ &= 4.0 k_B T \end{aligned} \tag{2.22}$$

given that $G_{3D} - \langle G_S \rangle = 10 k_B T$, which was chosen because of previous theoretical work[39] and is justified in that it corresponds to a non-specific dissociation constant $K_{ns} \approx 22 \mu M$, which is a typical value for non-specific protein-DNA affinity. In the correlated model with $\sigma_S = 1 k_B T$, the regime boundary rises to $8.9 k_B T$.

This treatment is correct within the parameters of our model and thus gives accurate results as regards the simulated results. When predicting a minimum k_f comparable with experimental data, it assumes that the off rate from the **S** mode to solution, τ_{1D}^{-1}, depends on no additional barrier beyond the free energy difference between the bound and unbound states. Considering the reverse reaction, this is equivalent to association being diffusion-limited, which is usually assumed. It further assumes that D_{1D} depends on the same τ_0 as does τ_{1D}, *i.e.* that the hydrodynamic friction in dissociating is comparable to rotating 40°, and that it is limited by disorder in the **S** mode rather than transition states between adjacent base pairs.

2.5.5 $K_{R/S}$ in the sequence-dependent model; point of transition between slow-sliding and fast-sliding regimes

Estimation of the equilibrium constant $K_{R/S}$ for the **S**→**R** transition requires deriving the quotient of the partition function over positions in the **R** state and the partition function over positions in the **S** state. Given our random energy model [39], these partition functions take the form:

$$Z = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(E-\mu)}{2\sigma^2}\right) \exp(-E) dE \tag{2.23}$$

This form does not give a good estimate of the partition function, however, because it assigns substantial weight to very-low energy states that are not expected to exist given the simulation genome length of 1000 bp. The lower bound of $-\infty$ must be replaced by a cutoff energy, E_c, equal to the expected value of the minimum value in the distribution, which, for M normally-distributed energies with mean μ and variance σ^2, is:

$$E_c = \mu - \sigma\sqrt{2\ln\frac{M}{\sqrt{2\pi}}} \tag{2.24}$$

With the lower limit of integration replaced with E_c, equation (2.23) evaluates to:

$$Z = \frac{\exp\left(\frac{\sigma^2}{2} - \mu\right)}{2}\left(1 - Erf\left(\frac{\sigma}{\sqrt{2}} - \sqrt{\ln\frac{M}{\sqrt{2\pi}}}\right)\right) \tag{2.25}$$

K_{eq} is thus

$$K_{R/S} = \frac{Z_R}{Z_S} = \exp\left(\frac{\sigma_R^2 - \sigma_S^2}{2} - \Delta G_{RS}\right)\left[\frac{1 - Erf\left(\frac{\sigma_R}{\sqrt{2}} - \sqrt{\ln\frac{M}{\sqrt{2\pi}}}\right)}{1 - Erf\left(\frac{\sigma_S}{\sqrt{2}} - \sqrt{\ln\frac{M}{\sqrt{2\pi}}}\right)}\right] \tag{2.26}$$

Setting $K_{R/S} = 1$ and solving for ΔG_{RS} will locate the transition between the fast-sliding ($K_{eq} \ll 1$) and slow-sliding ($K_{eq} \gg 1$) regimes. With $\sigma_R = 6.68 k_B T$ and $M = 1000$, the transition is found at $15.0 k_B T$ for $\sigma_S = 0 k_B T$ and at $14.5 k_B T$ for $\sigma_S = 1 k_B T$.

2.5.6 Optimal k_f given a constant k_u

We derive Equation 2.5 for the sequence-independent landscapes as follows. Using the approximate expression for $1/P_f$ (Equation 2.2), writing $k_{\text{res}} = \tau_{\text{res}}^{-1}$, and noting that $K_{R/S}$ must equal k_f/k_u, we have:

$$\bar{t}_s = \frac{M}{\bar{n}} \frac{k_f + k_{\text{res}}}{k_f} \big(\tau_{1D}(1 + \frac{k_f}{k_u}) + \tau_{3D}\big) \tag{2.27}$$

Partial differentiation with respect to k_f and setting the result equal to zero gives:

$$\frac{\partial \bar{t}_s}{\partial k_f} = \frac{\tau_{1D}}{k_u} + \frac{-k_{\text{res}}(\tau_{3D} + \tau_{1D})}{k_f^2} = 0 \tag{2.28}$$

and solving for k_f gives:

$$k_{f,opt} = \sqrt{\frac{k_{\text{res}}(\tau_{1D} + \tau_{3D})k_u}{\tau_{1D}}} \tag{2.29}$$

which is equivalent to Equation 2.5.

2.6 Outlook and Discussion

2.6.1 Optimal σ_S

The present work treated disorder in the **S** landscape as a binary, "on-off" parameter. But just as simulations were performed over a range of values for the other two chief parameters of interest, ΔG_{RS} and ΔG_{fold} (or k_f), and efficient and inefficient regimes were identified for them, the same could be done for σ_S. Especially when the **S** and **R** landscapes are correlated, an optimal σ_S can be expected to be found, so long as k_f is slow enough that pre-selection is necessary at all. Too low a σ_S would make pre-selection insufficiently strong to keep $1/P_f \approx 1$, while as σ_S approached σ_R, the utility of a sliding mode would vanish.

It is tempting to suggest, given the observation that most sequence-specific proteins whose D_{1D} on DNA has been measured slide with $\sigma = 1{\sim}2k_BT$ [42, 25, 22, 43, 44]

rather than with σ nearly zero, that many proteins do experience an effectively optimal σ_S. The diffusion coefficient of the sequence-independent C-terminal domain of tumor suppressor p53 indeed corresponds to a somewhat lower effective σ of $0.6k_BT$ [53]. On the other hand, that experimentally inferred σ_S falls within the likely range of the optimal may owe simply to the necessity of forming and breaking contacts between protein and DNA, and on a microscopic level correspond to barriers between base pairs rather than sequence-dependence in the protein-DNA complex's binding energy. A review of reported diffusion coefficients of sequence-dependent and sequence-independent DBPs as well as further experiments to measure diffusion coefficients, particularly of sequence-dependent DBPs on non-cognate DNA, would help address this question.

2.6.2 Beyond average search time for one particle

Our model is concerned with the mean search-and-recognition time for a single protein, and the physical parameters that allow that quantity to be minimized. The criterion for efficiency selected for by nature, however, may be some statistic other than the mean. For single-celled organisms competing with their neighbors for resources to grow and divide, the mean may indeed be the statistic selected for, but for multicellular organisms or other cooperative inter-cellular environments other statistics may be more important. In development, for instance, it may be more important to select against the probability that a cell would be especially laggard in executing some critical process, *i.e.* to minimize the rightward skew in t_s. In the case of severe heat or osmotic shock where survival is unlikely, it may be advantageous for a cell to attempt a "Hail Mary", and so maximizing the probability of an especially rapid search by a shock-response transcription factor or other DBP may be selected for.

In addition to implications of the evolutionary environment on the distribution of

protein-DNA search times, our model omits consideration of the fact that multiple copies of the same DBP may be active in a cell at once. Eukaryotes especially have a nuclear population of TFs and other DBPs in the range of $10^2 - 10^5$ [65]. They may thus be able to afford suboptimal parameters and inefficient searches. Indeed, measurements of the diffusive properties of human tumor suppressor p53 and its truncation mutants on DNA show that given a population of 500-5000 activated proteins in the nucleus, the protein requires an $\mathbf{S}\rightarrow\mathbf{R}$ transition rate of at least $\sim 700 s^{-1}$ absent kinetic pre-selection [53]. For comparison, *lac* repressor's transition rate is estimated to be $10^3 - 10^5 s^{-1}$ [55, 54].

Multiple copies of a transcription factor in the nucleus, as well as combinatorial gene regulation, change the statistics of the timing of gene expression. For a single transcription factor searching for a single target site, the distribution of search times should be approximately exponential, as the protein is performing trials (1D sliding rounds) with no memory until success. The distribution when multiple transcription factors are involved, however, should exhibit extreme-value statistics. In the case of many transcription factors searching for the same target site, as is generally the case in eukaryotes, non-combinatorially regulated transcription is activated by the *first* TF to reach the site. That is, the timing of activation is distributed according to the minimum value of a number of exponential random variables, which follows a Weibull distribution. For combinatorial regulation with only one copy of each necessary transcription factor, transcription is activated by the *last* TF to reach the site, and the timing of activation should follow a gamma distribution. For combinatorial regulation with multiplies copies of the transcription factors, the statistics becomes more complicated.

2.6.3 Spatial considerations

The two-mode model discussed here assumes that the protein slides on DNA without obstruction and that upon dissociation from DNA, it has an equal probability of reassociating at any site. Our group, however, has also studied models that consider the effect of "roadblocks", *e.g.*, nucleosomes [37], as well as systems wherein a protein has an enhanced probability of reassociating close to where it left the DNA [38, 37]. These studies have assumed, however, a flat landscape and instant recognition by a protein upon reaching its target site. Unifying the two-mode model with the models used in these studies would offer a more holistic picture of protein-DNA search and have implications not found in either of model alone.

For instance, if obstacles are dense enough around the target site, they would limit $\bar{n}$ to a smaller value than that given in Equation 1.4. This would have the effect of making searches more redundant, reducing $\bar{n}$ and increasing τ_{res} without having any effect on τ_{1D}. The increase in τ_{res} would allow for a more relaxed lower limit on k_f for efficient folding. A protein that executes its searches *in vivo* on crowded DNA and indeed folds with $1/P_f \approx 1$ may spuriously appear in experiments using naked DNA to have a k_f in the slow-folding regime.

The presence of obstacles would also bear on the optimum σ_S for the protein should it employ kinetic pre-selection. One of the ways in which σ_S affects $\bar{t}_s$ is through an $\exp(-\frac{7}{4}\sigma_S^2)$ dependence of D_{1D}. More ruggedness in the **S** state corresponds to a smaller diffusion coefficient and thus fewer sites sampled per round of 1D sliding, *i.e.* a smaller $\bar{n}$. If $\bar{n}$, however, is limited by obstacles rather than by D_{1D}, then larger values of σ_S have less of a tardative effect on the search time and the optimal σ_S will be greater than otherwise. In the extreme case, where the optimal σ_S approaches σ_R, a distinct **S** mode no longer contributes to an accelerated search.

2.6.4 Kinetic proofreading and enzymatic reactions

Although kinetic pre-selection is not a traditional proofreading mechanism, it overlaps with "kinetic proofreading" [66] in its function. In kinetic proofreading, a biochemical system makes use of an intermediate, metastable complex preceding an irreversible enzymatic step. Substrates that form complexes that are much shorter-lived than the time for the reaction dissociate from the enzyme before the reaction can proceed, while substrates that form more stable complexes survive long enough to undergo the reaction. Although this study was motivated chiefly by the speed-stability paradox, which concerned proteins that needed to bind persistently to a specific site, the kinetic pre-selection mechanism can also account for the observed speed and *specificity* of DNA-binding proteins with enzymatic activity on DNA, such as oxoguanine glycosylase, which irreversibly cleaves the glycosidic bond upon recognizing an 8-oxoguanine lesion.

2.7 Acknowledgements

J.L. acknowledges support from the National Science Foundation.

Part II

Kinetics of p53's diffusion on DNA

Chapter 3

Introduction

As discussed in Part I, the experimentally observed and biologically necessary speed with which DNA-binding proteins (DPBs) reach their target sites and bind them stably can be explained by a two-mode model of DBP-DNA interactions. An important class of DNA-binding proteins is transcription factors (TFs). Transcription factors are proteins that bind DNA at specific sites or motifs, and activate or repress transcription of particular target genes. The proper timing of gene expression requires that TFs efficiently locate and bind their target sites within a genome.

A eukaryotic transcription factor faces a number of challenges in finding its target site or sites. In bacteria, TF genes are often co-localized with their binding sites, and the coupling of transcription and translation causes bacterial TFs to be synthesized near to their targets [67]. Eukaryotic TFs, however, are translated in the cytosol and thus enjoy no such "head start". Furthermore, upon induction into the nucleus, TFs locate their target, or *cognate*, sites entirely by passive transport. Lastly, theoretical and experimental studies [19, 21, 61, 42] have shown that transcription factors, and sequence specific DBPs generally, have affinity to the vast excess of accessible non-cognate DNA (10^7–10^9 bp).

As the motivation for the theory developed in Part I is to explain how DBPs including eukaryotic TFs can, despite these challenges, efficiently search for and recognize their targets, it is fitting to test the theory by assessing its compatibility with experimental data on TFs undergoing diffusive search on DNA. As will be seen, a number of ensemble-averaging experiments have been performed that offer indirect evidence for a 1D/3D search mechanism, although these experiments do not address the question of whether TFs bind DNA with multiple modes, and are limited in their ability to address other details of TF-DNA interactions.

This chapter of the thesis reviews the history of experiments on DBPs undergoing 1D diffusion on DNA, lays out the need for single-molecule experiments, in particular on eukaryotic TFs, and then discusses the single molecule techinques used in our experimental work and the eukaryotic TF studied, the tumor suppressor p53. The following Chapters (4 and 5) in this Part present our research on the sliding kinetics of p53, and the final Chapter (6) outlines some of the challenges in studying molecular search processes and suggests future experiments, both those relevant to the general problem and those that might shed light specifically on p53.

3.1 Experimental studies of 1D diffusion of proteins on DNA

3.1.1 Ensemble-averaging experiments

The first directly controlled experiment supporting 1D/3D facilitated diffusion as a means of target localization by DBPs for their target sites on DNA, by Jack *et al.* [20], was discussed in Chapter 1. The researchers found that the restriction enzyme EcoRI cut its target site faster when the site was flanked by longer sequences of non-cognate DNA. Also discussed in Chapter 1 in greater detail was a subsequent study demonstrating interchange

between DNA molecules of the restriction enzyme EcoRV, by Halford and co-workers [21], evidenced by a scheme involving placing the restriction site in various places in minicircle-plasmid catomers and inferring transfer from one circle to the other.

One-dimensional search can proceed in theory either by *sliding*, in which proteins maintain constant contact with the DNA, or by *hopping*, in which they make a local excursion away from the DNA but reassociate nearby (∼1 persistence length). To distinguish these two modes of translocation, Gowers *et al.* [21] studied the effects of the orientation and spacing of two nearby target sites of the restriction enzyme BbvCI. They found that the sites,when spaced by less than 50 bp, were more efficiently cleaved when oriented in the same way than when oriented in opposite directions, but that orientation made no difference when the sites were spaced by greater than 50 bp. From this they concluded that the sliding mode was significant on lengthscales below 50 bp for BbvCI, but that above 50 bp, hopping and/or long-range jumping dominated.

The ensemble-averaging studies discussed thus far used evidence of the product of an enzymatic reaction as a readout. Transcription factors lack such biochemical activity, and so studying them requires different experimental techniques. The transcription factor p53, the subject of this Part of the thesis, was suggested to slide on DNA by measuring the rate at which it dissociated from DNA bearing a target site [68]. They found that the dissociation rate of a p53-oligonucleotide complex decreased upon blocking the ends of the oligonucleoide with stretpavidin proteins, which implied a greater off-rate from the ends of the DNA and the ability of p53 to translocate from its binding site in the center of the oligo to the ends so as to take advantage of this faster "escape route".

The lack of biochemical or other ready readouts from TF–cognate site recognition, as opposed to binding between a TF and whatever DNA construct is employed in an assay, prevents the use of techniques such as those discussed earlier that were used with restriction

enzymes in elucidating those proteins' recognition kinetics and translocational dynamics. To examine these properties of TF-DNA interactions, a more direct technique that can resolve individual TF's binding to, moving along, and dissociating from DNA is required. Such single-molecule techniques additionally afford the ability to study the heterogeneity among a population of molecules. This is particularly important for phenomena such as gene expression and DNA repair, the proper functioning of which is governed by extreme-value statistics, or other quantities that are difficult to investigate by bulk biochemical methods whose readout is an ensemble average. Whether the transcription of RNA for cell-cycle-arrest proteins or whether DNA-damage repair, for example, is effective in preventing mutagenesis can depend on the time it takes for the fastest individual transcription factor (or the fastest few in the case of combinatorial transcriptional regulation) or repair enzyme out of a population of hundreds or thousands to find and recognize their targets rather than on the population average.

3.1.2 Single-molecule experiments

Investigations of the 1D-diffusional properties of sequence-specific DNA-binding proteins have been discussed in Chapter 1. These studies have taken place within the past half-decade; earlier single-molecule microscopy experiments examined the sequence-independent *E. coli* RNA polymerase diffusing on DNA [69]. The researchers imaged the protein's movement using total internal reflection fluorescence microscopy (TIRFM), which we also use in the work described in Chapters 4 and 5.

More detailed data on mechanisms of polymerase diffusion on DNA was provided by Kim *et al.* [70]. In an experimental setup very similar to the one used in this thesis's work, T7 RNA polymerase was imaged in a flow cell in which DNA was bound and stretched by laminar flow. Based on the idea that non-specific polymerase-DNA affinity was largely

electrostatic, they measured the protein's diffusivity at a range of solution ionic strengths. If the protein diffuses by hopping, then a higher salt concentration will decrease the rate of reassociation of protein to DNA after a momentary dissociation, and thus the protein will spend more time diffusing in 3D (but still near the DNA), where it experiences a larger diffusion coefficient than it does when on DNA. If the protein maintains continuous contact with the DNA, however, then the 1D diffusion coefficient should be independent of the solution's ionic strength [71, 72, 25]. The researchers found that T7 RNA polymerase's 1D diffusion coefficient was indeed independent of salt concentration, and concluded that it translocates along DNA by sliding rather than by hopping.

Coupling of rotational and translational diffusion

The distinction between hopping and sliding is a critical one in testing the validity of our proposed two-mode model. The model relies on proteins sampling the sequence information of DNA while diffusing along it, and thus requires that proteins slide rather than hop in order to take advantage of the facilitated search it offers, including kinetic pre-selection. A protein that slides along DNA while sampling the sequence information is expected to follow the helical pitch of DNA to maintain the structure of the protein-DNA complex along the DNA. This coupling of rotational diffusion with linear diffusion results in much greater (10^2-10^3 times) hydrodyamic drag experienced by proteins [62] relative to drag from purely translational movement.

The Stokes-Einstein relation gives the diffusion coefficient D_{1D} for a translating particle as:

$$D_{1D} = \frac{k_B T}{6\pi\eta a} \tag{3.1}$$

where η is the viscosity of the solvent and a is the hydrodynamic radius of the particle. If

a particle must rotate as it translates, the diffusion coefficient becomes [62]:

$$D_{rot,1D} = \frac{k_B T}{6\pi\eta a\left[1 + (4/3)(2\pi)^2(a/\tau)\right]^2} \tag{3.2}$$

where the expanded term in the denominator accounts for additional friction due to rotation. τ is the axial length of a helical turn, which is 3.6nm in the case of DNA. The diffusion coefficients given in Equations 3.2 and 3.1 are the upper limits of diffusion coefficients for particles—they assume no additional energy barriers between positions on DNA.

Identifying whether a protein's linear diffusion along DNA is coupled to rotational diffusion is necessary to determine the height of energy barriers between positions. A small diffusion coefficient for a protein sliding on DNA could owe to large barriers, or it could owe to having to track the pitch of the helix. Recently, Kochaniak *et al.* performed single-molecule microscopy experiments on proliferating cell nuclear antigen (PCNA), which is known to take the form of a ring around DNA, and to which various other DNA-reading or -modifying proteins can attach. By measuring the protein's diffusion coefficient as a function of solution viscosity and protein size, they determined that PCNA diffuses too quickly along DNA to track the helix 100% of the time; rather, it spends most of its time tracking the helix and a small share of its time "slipping" [73].

Unlike with PCNA, which has no need to read the DNA sequence along which it slides, we have a strong *a priori* reason to believe that sequence-specific DBPs such as repair enzymes and transcription factors must track the DNA helix. For their sliding to be efficient as a means of accelerating target search, the barriers between adjacent base pairs must be $\lesssim$ 1–2 $k_B T$, as discussed in Part I. Indeed, nearly all sequence-specific DBPs that have been shown to diffuse on DNA do so with diffusion coefficients corresponding to helical tracking with the requisitely small inter-base-pair energy barriers, as determined using Equation 3.2 [74, 23, 75, 76], a result that, given the range in sizes of proteins studied, is unlikely

to arise from non-rotating diffusion across much larger energy barriers that happens to yield diffusion coefficients all just below the proteins' theoretical maximums for rotational diffusion.

Single-molecule experiments have been crucial to understanding the 1D-diffusive behavior of DBPs on DNA. In the following section, we discuss the techniques used in our experimental work.

3.2 Single-molecule techniques for studying protein diffusion on DNA

3.2.1 TIRFM

Optical imaging of any object, from a baseball to a protein, requires that the object be distinguishable from the rest of the field of view, and that the object be not move too rapidly out of the focal depth and field of view, given the frame rate of the imaging device. Total internal fluorescence optical microscopy (TIRFM) addresses the first of these requirements, while our immobilization of DNA in a flow cell addresses the second (discussed in Section 3.2.2).

TIRFM allows individual fluorescent molecules to be imaged by selectively illuminating a thin layer of sample (100 ~200 nm) at a sample-glass interface (Figure 3.1). This is achieved by directing an excitation laser at the interface at an oblique angle greater than a certain *critical angle*, which is a function of the refractive indices of the sample (typically an aqueous buffer, with $n = 1.33$ to 1.38) and the glass ($n = 1.5$). At or above the critical angle, the intensity of light propagating through the interface is zero, but a non-propagating evanescent field exists, the intensity of which decays exponentially with increasing distance from the interface with decay constant d, also called the penetration depth, which is given

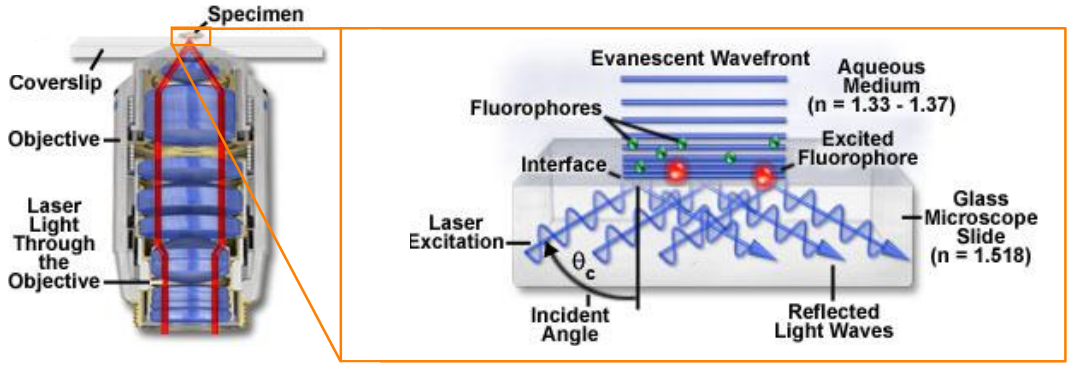


Figure 3.1: Total internal reflection fluorescence microscopy (TIRFM) implementation. Excitation laser strikes the interface between an aqueous sample and a glass cover slip. No light is propagated through the sample; rather, an evanescent field arises with intensity decaying exponentially away from the interface. The beam of light is totally reflected, creating a thin evanescent field of excitation energy in the aqueous medium. The evanescent field intensity decays exponentially with increasing distance from the interface. Fluorophores within 100 ~200 nm of the interface are excited; the great bulk within the solution are not. Figure adapted from the Nikon "MicroscopyU" website.

as:

$$d = \frac{\lambda}{4\pi(n_{\text{glass}}^2 \sin^2(\theta) - n_{\text{sample}}^2)^{1/2}} \tag{3.3}$$

where λ is the wavelength of the light and θ is the angle of incidence. Since $\sin^2(\theta)$ is always ≤ 1, n_{sample} must be less than n_{glass} to obtain a real penetration depth, a limitation not shared by the older technique of confocal microscopy[1]. A great advantage of TIRFM, however, is that the depth of the sample illuminated and concommitant background fluorescence is approximately an order of magnitude less. The non-illumination of the 10^2 to 10^4 times as many fluorescent particles in the sample solution as within the penetration depth allows individual distinct particles to be imaged.

TIRFM is typically implemented either using an objective lens to collect photons

[1]For values of n_{sample}, n_{glass}, and θ where d would be imaginary, no evanescent wave is formed and the intensity of propagating transmitted light is nonzero.

from the sample and a separate focusing lens on the other side of the sample through which the excitation laser is directed, or using the same lens for both focusing and angling the excitation light as well as collecting photons from the sample. We employ these second of these two setups, illustrated in Figure 3.1. With this setup, an emission filter is necessary to exclude light of the excitation wavelength but pass light of the emission wavelength(s).

3.2.2 Flow-cell assay

The p53 and lambda-phage DNA we wish to image would, if free in solution, diffuse out of the 100 ~200 nm illuminated by the evanescent field too quickly to be imaged. To confine our molecules of interest to this region, we employ a flow-stretching technique (Figure 3.2), which is simpler than optical and magnetic trapping. DNA bearing one element (typically the small molecule) of a stable linking pair such as biotin-streptavidin or digoxigenin–anti-digoxigenin is flowed into the cell and attaches to the cover slip, which has been functionalized using the other member of the pair [61, 43]. Other methods, such as a "DNA curtain" anchored in a lipid bilayer, have also been employed [77]. The drag force of the flowing buffer keeps the DNA stretched and near the surface, although its fluctuations are not negligible and, for the work in Chapter 5, require measurements of its dynamics so that they may be separated from the dynamics of proteins bound to the DNA. After the DNA has bound to the surface, protein is flowed in, illuminated, and imaged.

3.2.3 Imaging considerations

Measurements of any phenomenon require excluding extraneous signals from drowning out the signals from the phenomenon of interest while at the same time ensuring that the desired signals are strong enough to be detected. In single-molecule fluorescence microscopy, the former amounts to keeping background fluorescence at an acceptably low level,

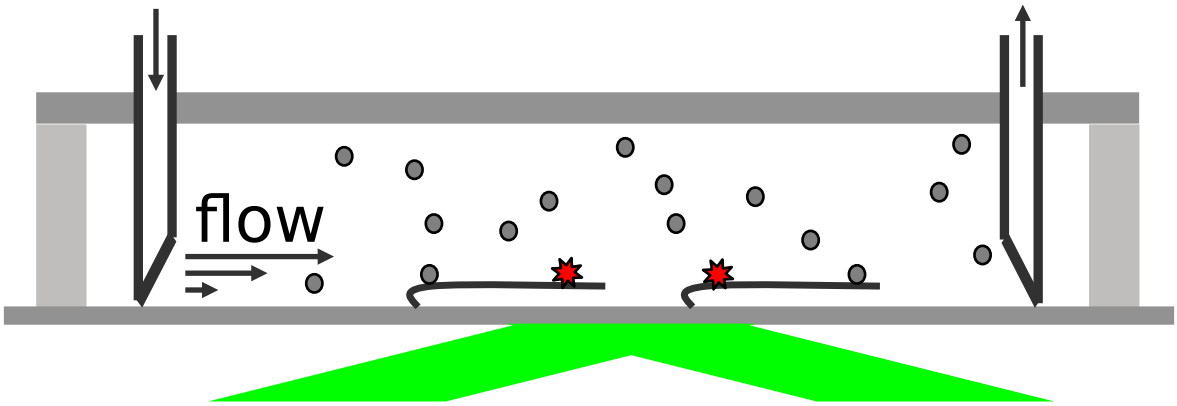


Figure 3.2: Illustration of flow cell. The flow-stretching technique subjects DNA bound to the surface of a cover slip to laminar flow, which stretches out the DNA and keeps it close to the cover slip. Using TIRFM, only fluorescent molecules on or near the surface are excited.

and the latter means achieving a strong and robust signal from particles under study, which is limited by the detector and by the fluorescent particles themselves.

Background fluorescence

As discussed in Section 3.2.1, TIRFM restricts illumination of the sample volume to the 100 ~200 nm nearest to the glass-buffer interface. As the evanescent field does not drop in intensity to zero beyond this point, however, the concentration of fluorophores in the buffer must still be limited. Generally, sub-nanomolar concentrations of fluorophores are required to keep background fluorescence low enough in TIRFM. In addition to fluorescence from bulk solution fluorophores, another potential source of background in our TIRFM setup, wherein the excitation light is focused onto the sample with the same lens as is used for collection, is the failure to adequately filter out excitation light with a well-chosen band-pass filter between the objective lens and the camera. A source of background that cannot be filtered out, however, is Raman scattering, although its effect is usually small compared to background fluorescence except at high excitation levels with low concentrations of fluorophores.

Camera and statistical noise

The magnitude of the signal from fluorophores in the sample detected by a camera is equal to the product of the number of incident photons, P, and the quantum efficiency, Q, which is the proportion of incident photons that are detected by the camera. Contributions to noise include, in the case of a CCD camera such as the one we used on our experiments, (1) readout noise, (2) dark current noise, and (3) shot noise.

1. Readout noise, δ_r, arises from the processes involved in amplifying and converting the photoelectrons created by incident light into voltages. The magnitude of readout noise can be decreased by decreasing the imaging frame rate.

2. Dark current noise, δ_d, owes to thermally generated photoelectrons. It is to reduce δ_d that CCD cameras are cooled, typically to -25 or $-65\,^{\circ}C$.

3. Shot noise, δ_s, is inherent to the Poisson distribution chracterizing the number of incident photons. It is simply the standard deviation of the mean number of detected photons, that is, $\sqrt{QP}$.

The total noise is sum in quadrature of the component noise contributions, and so the signal-to-noise ratio is:

$$SNR = \frac{QP}{\sqrt{\delta_r^2 + \delta_d^2 + QP}} \tag{3.4}$$

Before the development of the electron-multiplying CCD (EM-CCD), readout noise was the dominant source of noise in single-molecule imaging. The basic principle of an EM-CCD camera is similar to that of an avalanche photodiode, in which photoelectrons generated by incident light stimulate high-potential "store" electrons to pass down a voltage gradient, producing a gain in signal prior to its transmission to the readout circuitry. The

gain is great enough for readout noise to have been reduced below noise generated by Poissonian fluctuations in incident photons.

Fluorphore characteristics

In biological single-molecule microscopy experiments, desirable optical properties of fluorophores include minimal photobleaching, a high extinction coefficient, and high quantum yield. Photobleaching usually involves a chemical reaction between an excited fluorophore and some other molecule, or between a fluorophore and a reactive oxygen species. The probability of the former kind of reaction depends, among other factors, on the excitation lifetime of the fluorophore and its specific chemical properties. The latter source of photobleaching can be combatted using an oxygen-scavening system and/or reagents that catalyze the decay of singlet oxygen to the less reactive triplet state.

The extinction coefficient of a fluorophore, or any molecule for that matter, is its probability at a given wavelength to absorb a photon. Small extinction coefficients require more intense excitation light to produce the same signal, which has the undesirable effect of increasing background due to scattering or autofluorescence. A fluorophore's quantum yield is the ratio of emitted to absorbed photons. Nonradiative decay is responsible for quantum yields less than unity, modes of which include relaxation through a triplet state, dissipation of energy into vibrational modes, or quenching through interaction with another molecule, often dioxygen, or a moiety of the biomolecule to which it is coupled.

Additional properties of superior fluorophores include non-interaction with the biological system, solubility, and the absence of steric, hydrodynamic, electrostatic or other artifacts relevant to the measurements.

Position uncertainty

Diffraction limits the resolution of nearby sources of light, but in single-molecule imaging, the position of an isolated fluorophore can determined with a precision well below the diffraction limit. The point-spread of the system is a two-dimensional Gaussian distribution that is effectively binned by camera pixels. Sub-diffraction and sub-pixel position determination usually proceeds by fitting this intensity distribution to a two-dimensional Gaussian, the peak of which is recorded at the particle position.

The uncertainty in of position in an arbitrary direction is given by:

$$\delta = \sqrt{\frac{s^2}{N} + \frac{a^2/12}{N} + \frac{8\pi s^4 b^2}{a^2 N^2}}, \tag{3.5}$$

where s is the standard deviation of the microscope point-spread function, N is the number of photons collected (equal to QP above), a is the pixel linear size, and b is the standard deviation of the background fluoresence intensity [78].

3.2.4 Drift due to flow

The flow-stretching technique relies on a drag force exerted on DNA to keep it extended in a linear conformation. While this eases the recording and analysis of trajectories of particles on it, it introduces a bias in particles' random walks. The work in Chapter 4 determines an aggregate drift velocity and then subtracts this drift component from all trajectories (see Chapter 4, *Methods*). The resulting trajectories' plots of their mean-squared displacement versus time-window appear linear, as expected for normal diffusion, and the slope of the MSD-versus-time-window plots are used to determine the diffusion coefficients of the particles.

The work in Chapter 5, however, is concerned with proteins' diffusion coefficients as a function of their position along the contour of the DNA. Modifying their trajectories

as is done in Chapter 4 would result in the misassignment of their positions. Rather, we consider the diffusion coefficient D and the drift velocity v contributing to each movement of a particle as parameters to calculate using maximum likelihood estimation (MLE) (see Chapter 5, *Methods*).

3.2.5 DNA fluctuations

As mentioned in Section 3.2.2, the DNA itself undergoes Brownian motion, in both the longitudinal (direction of the flow) and transverse (perpendicular to the flow direction) dimensions. It is the longitudinal fluctuations that affect the estimation of protein diffusion coefficients, and these effects are large enough to require careful consideration when analyzing data. A particle that is immobile on the contour of the DNA will appear to have nonzero MSD for time-windows shorter than the timescale of the DNA fluctuations. Since the DNA is bound to the surface of the flow cell, its Brownian motion is bounded, and for sufficiently long time-windows, the immobile particle's MSD will cease to increase with increasing time-window duration. The lengths of trajectories are such that in Chapter 4, we were able simply to omit the first few time-windows from our fits of the MSD plots, and consider only displacements that took place over sufficiently long time.

Such an approach toward the DNA fluctuations, however, was found to require excluding unacceptably much data for the work in Chapter 5. Instead we developed a unified MLE treatment that separated apparent diffusion owing to DNA from diffusion owing to proteins undergoing unbiased random walks on the DNA contour and from the observed bias from drift due to flow (see Chapter 5, *Methods*). This procedure required measurements of the DNA fluctuations using fluorescent probes bound covalently to the DNA, as well as Brownian dynamics theory of a tethered polymer in shear flow [79].

3.3 Tumor suppressor p53

Single-molecule experiments that have assayed whether site-specific DBPs hop or slide have been performed on a variety of prokaryotic proteins, but until the work described in Chapter 4, none had been performed to our knowledge on eukaryotic site-specific DBPs[2]. We decided to investigate the 1D diffusional properties the human tumor-suppressor transcription factor p53 for a number of reasons. Studying eukaryotic TFs in this capacity was novel. Even though a variety of prokaryotic DBPs had been shown to slide on DNA, it was thought possible that the presence of nucleosomes in eukaryotic DNA as well as the often much-larger cellular population of eukaryotic DBPs might make sliding less useful and less necessary in eukaryotes. The tumor suppressor p53 was particularly promising for study, as an earlier experiment had given indirect evidence that it slid on DNA, and that a specific domain, the C-terminal domain, was responsible for its sliding capability [68]. The suggested division of recognition and sliding functionalities into distinct domains offered a straightforward way to study them separately or together and thereby test the two-mode model discussed in Part I, by the use of truncation mutants of the full-length protein. Lastly, p53 is a protein of great medical importance.

3.3.1 p53's function and structure

The eukaryotic tumor suppressor p53 is known as the "guardian of the genome". In response to DNA damage and other oncogenic stress, p53 is activated and induces the transcription of genes that, depending on the cell cycle and the extent of the damage, can instigate DNA damage repair, cell-cycle arrest, senescence, or apoptosis [81, 82, 83]

[2]Excepting perhaps human oxoguanine DNA glycosylase, hOgg1 [61]. This protein, however, is homologous to the bacterial protein with the same function, MutM, and is active on mitochondrial DNA rather than in the eukaryotic nucleus [80].

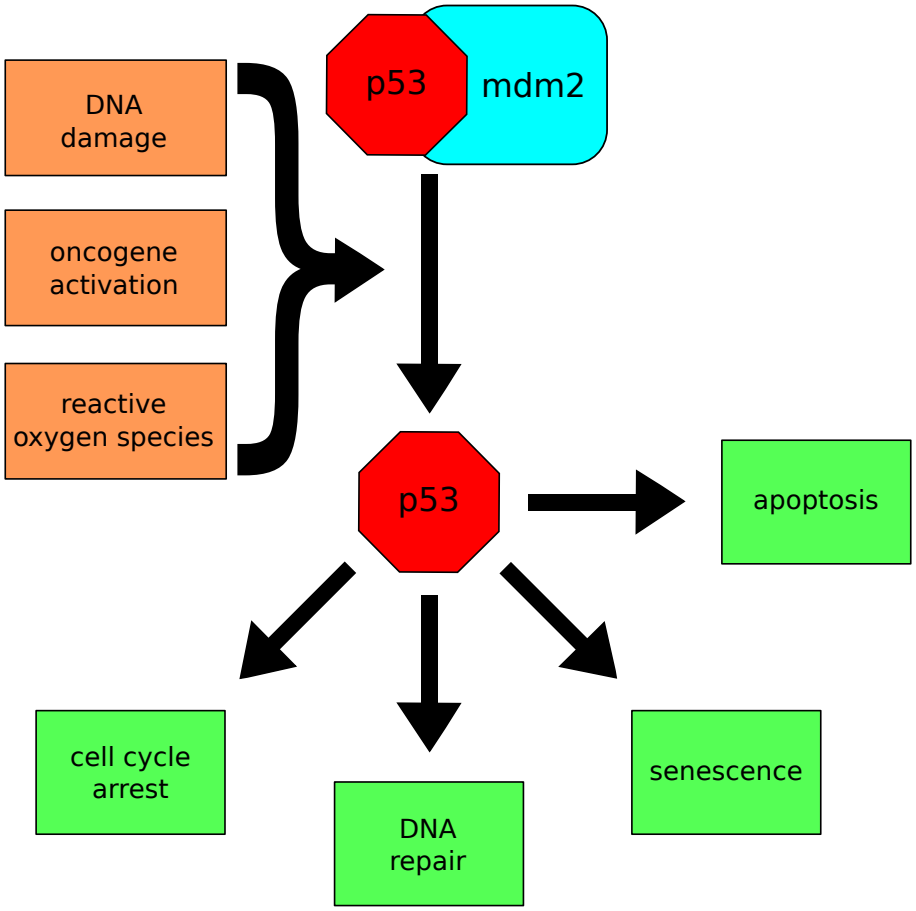


Figure 3.3: p53 suppresses tumorigenesis in response to threats to genome integrity. The transcription factor is activated by the disruption of its interaction with mdm2, whereupon it initiates transcriptional programs that avert oncogenesis.

(Figure 3.3). The protein's importance in preventing oncogenesis is underscored by the finding that more than 50% of all human cancers sequenced have a mutation in the gene for p53 [83]. For p53 to be effective in preventing tumorigenesis, it must reach its target genes quickly enough in response to activation to prevent the replication of damaged DNA or mitosis. This need for speedy location of and binding to its target sites makes it an attractive candidate to test the two-mode model.

p53 consists of four distinct domains, with their own distinct functions: an N-terminal regulatory domain, the core sequence-specific DNA-binding domain, a tetramer-

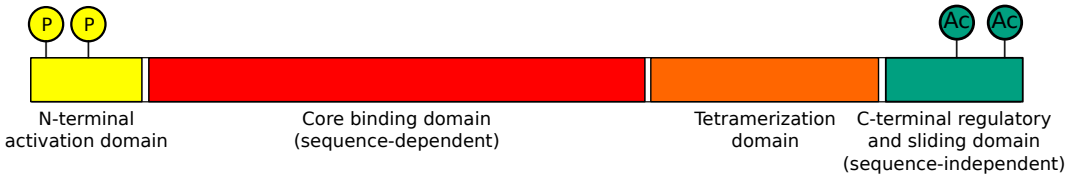


Figure 3.4: The domains and selected post-translational modifications of p53. Starting from the N-terminus, the protein consists of an N-terminal regulatory domain that plays the chief role in p53 activation, by being phosphorylated at a number of residues. The bulk of the protein consists of the core DNA-binding domain, which has a sequence-dependent binding affinity. The core domain is also responsible for dimerization of monomers, while the tetramerization domain is necessary for the dimerization of dimers. The lysine-rich C-terminal domain is unstructured in solution, binds non-specifically to DNA, and has been implicated in previous work [68] to be responsible for p53's sliding modality, which was later shown more definitively by our groups [53]. Color scheme follows the 3D cartoon representation of p53 in Figure 1.3, (save for the N-terminal domain).

ization domain, and a C-terminal domain whose role has only recently become largely understood (Figure 3.4). Activation of the protein, in response to irradiation, DNA damage, or other potentially genome-destabilizing stress, entails phosphorylation of its N-terminal domain [84, 85]. This disrupts the interaction between p53 and its negative regulator mdm2. Depending on a number of factors, p53 then proceeds to activate various pathways to ensure genome stability (Figure 3.3).

The core domain of p53 is responsible for the recognition of its response elements (REs) [86, 87]. It binds weakly to non-cognate DNA [88] and strongly to cognate sequences, of which hundreds have been identified (discussed further in Section 3.3.2). Owing to the substantial degeneracy of its binding motif, sufficiently long non-repetitive DNA, such as lambda-phage DNA used in previous single-molecule experiments as well as the present work, can be expected simply by chance to include close matches to p53 cognate sequences.

Unlike most TFs, p53 has a second DNA-binding domain—the positively-charged C-terminal domain. Initial studies suggested that the domain was a negative regulator on the core domain. In *in vitro* studies, Hupp *et al.* found that deleting the C-terimanl

domain, phosphorylating or acetylating it (both of which reduce the positive charge), or subjecting it to an antibody increased the affinity of p53 toward specific sites on short oligonucleotides [89].

Subsequent experiments suggested a positive regulatory role, however, for the C-terminal domain. As described in Section 3.1, the domain was suggested to be necessary for one-dimensional translocation of the protein on DNA, based on measurements of p53's dissociation rate from DNA in which the ends of an oligonucleotide on which the protein was incubated were either free or blocked, and thus the escape mechanism of sliding to the ends of the DNA was or was not available [90]. The same researchers also demonstrated that the ΔC-terminal mutant was slower *in vivo* to transactivate its targets.

These observations can be reconciled by the application of the two-mode model discussed in Part I. If p53's core DNA-binding domain is responsible for **R**-mode binding while the C-terminal domain is responsible for **S**-mode binding, then the C-terminal domain should accelerate target recognition and thus lead to more rapid transactivation, as observed by McKinney *et al.* Disabling or removing the C-terminal domain, however, stabilizes binding in the **R** mode to cognate sites. If the two-mode model is indeed applicable to p53, then it the protein should be able to slide efficently with $\sigma \lesssim 1 - 2k_BT$, and it should also exhibit sequence-dependent sliding kinetics. We find that both predictions are borne out, the former in experiments discussed in Chapter 4, and the latter in Chapter 5.

3.3.2 Special considerations for p53

p53 is a dimer of dimers

As described in Section 3.3.1, p53 is a homotetramer consisting of two dimers bound through the tetramerization domain. While the active species *in vivo* is tetrameric and binds canonically to 20-bp response elements, dimeric p53 has been observed to bind

to 10-bp half-sites[91]. Additionally, binding of tetrameric p53 to half-sites has recently been shown to play a role in transcriptional activation at high p53 expression levels[92]. We hypothesized that p53 might be able to bind DNA in a hybrid, "hemi-specific" mode, with one dimer binding in the **R** mode and the other in the **S** mode. While no structural studies have been performed on p53 bound to a half-site, crystallographic measurements on the restriction enzyme BstYI[50] provide direct evidence of at least one protein bound hemi-specifically.

If p53 can bind hemi-specifically, then it should encounter a site of enhanced affinity (*i.e.* a half-site) on average once out of a number of base pairs equal to 2 raised to the logo information content in bits. The canonical binding motif for p53 is a 20-bp sequence consisting of two 5'-RRRGWWCYYY-3' half-sites, with a gap of 0–14 bp. Hundreds of p53 response elements (REs) have been identified [93, 94, 95, 96, 97, 98], most of which stray in at least one position from the canonical motif. As will be discussed in Chapter 5, position weight-matrices built from a catalogue of known binding sites have only ∼5 bits of information per half-site.

With scarcely more than 5 bits in its half-site sequence logo, p53 is expected to have a half-site every $\sim 10^{1.5}$ bp. The protein indeed binds these half-sites with a lower K_d than it does to random DNA sequences, according to *in vitro* measurements of the affinity of p53 for oligonucleotides bearing random DNA, half-sites, and full-sites [88]. Inferences from measurements of the protein's diffusivity on lambda-phage DNA will have to take into account the possibility, then, of a substantial share of sites favoring binding half in the **S** mode and half in the **R** mode.

The two p53 dimers have been found by Weinberg *et al.* [88] to exhibit cooperativity in their binding to response elements. If the binding of one dimer did not influence the binding of the other dimer, then the enhancement in affinity the protein experiences

toward a full-site relative to random DNA should be simply the product of the enhancement for the left half-site times the enhancement for the right half-site (Equation 3.6). Instead, the researchers found that the tetramer binds more strongly to a full-site than would be predicted based only on its affinity toward the separate halfsites, with the two dimers interacting cooperatively with a Hill coefficient of 1.8.

$$\begin{aligned}
\text{No cooperativity:} \quad & \frac{K_d(\text{left half-site})}{K_d(\text{random DNA})} \times \frac{K_d(\text{right half-site})}{K_d(\text{random DNA})} = \frac{K_d(\text{full-site})}{K_d(\text{random DNA})} \\
\text{Cooperativity:} \quad & \frac{K_d(\text{left half-site})}{K_d(\text{random DNA})} \times \frac{K_d(\text{right half-site})}{K_d(\text{random DNA})} > \frac{K_d(\text{full-site})}{K_d(\text{random DNA})} \\
\text{Anti-cooperativity:} \quad & \frac{K_d(\text{left half-site})}{K_d(\text{random DNA})} \times \frac{K_d(\text{right half-site})}{K_d(\text{random DNA})} < \frac{K_d(\text{full-site})}{K_d(\text{random DNA})}
\end{aligned} \tag{3.6}$$

The cooperativity in binding to full-sites affects the energy landscape the protein should experience, making fully-specific (both dimers in the **R** mode) binding to full-sites more imporant relative to hemi-specific binding than it would otherwise be in contributing to any reductions in the protein's diffusivity when sliding on DNA.

Experimental challenges

Wild-type p53 at physiological temperature has a stability of only 2–3 kcal/mol [99], and at room experimental temperature of 6 kcal/mol [100]. Fortunately, an functionally identical mutant with enhanced stability has been engineered that allows for experiments on p53 of longer duration, before the protein misfolds or aggregates [99], and this mutant was available for our study.

Another potential concern when working with p53 is that concentrations low enough for single-molecule microscopy are lower than the dimer-tetramer K_d of 20 nM [101]. Fortunately, the half-life of the tetramer, which is the active species in *in vivo* transcrip-

tional activation, is approximately 3 hours [101] at room temperature. Still, to ensure that measurements of p53's diffusivity on DNA are performed on a homogeneous population of tetramers, fresh aliquots of p53 are needed frequently when collecting data. After dilution to sub-nanomolar concentrations, 5% of tetramers dissociate within 15 minutes, for instance.

* * *

In the following chapters, I discuss studies of the tumor-suppressing transcription factor p53's one-dimenstional diffusivie properties on flow-stretched lambda-phage DNA. In Chapter 4, I present our measurements of the diffusion coefficient of p53 on lambda DNA without regard to position on the DNA, and addresses whether the protein slides or hops, and in the event of the former, whether it tracks the helix while sliding. These questions bear on the applicability of the two-mode model presented in Part I to p53. The work of Chapter 5 assesses the functionality of p53's sliding on DNA by identifying whether its sliding kinetics are sequence-dependent, as predicted by the two-mode model. It further explores the importance of hemi-specific binding in p53's interaction with DNA. Finally, Chapter 6 discusses the implications of the work in the preceding chapters, both by themselves and united with the findings of Part I.

Chapter 4

Aggregate diffusional properties of p53 on DNA

In this chapter, I discuss experiments that lay the foundation for subsequent work in demonstrating that the two-mode model put forth in Part I accurately describes p53's sliding kinetics on DNA. These experiments demonstrate, most basically, that p53 indeed undergoes 1D-diffusion on DNA. Furthermore, p53 maintains contact with DNA while it diffuses, which is a necessary condition for the two-mode model. Lastly, its measured diffusion coefficient implies that it experiences a smooth enough landscape to benefit from facilitated diffusion.

4.1 Introduction

The tumor suppressor p53 is a transcription factor that responds to stresses, such as DNA damage, oxidative stress, heat shock, and deregulated oncogene expression, by inducing cell-cycle arrest or apoptosis [102]. The protein binds non-specific DNA through its highly basic C-terminus domain [103] and can undergo one-dimensional (1D) diffusion on DNA using

this domain [104]. This 1D diffusion has been suggested both to regulate negatively and positively gene activation. Experiments that have examined the dissociation of p53 from short DNA have shown that deleting the C-terminus [104, 103, 105, 106, 107] or replacing it with the neutral C-terminus of the related p73 protein [105] slows the dissociation of p53 from its promoter. Moreover, for wild-type p53, blocking the ends of the DNA [90], circularizing the DNA [104], or increasing the length of the DNA [105, 106] slow the rate of dissociation, suggesting that p53 relies on 1D-diffusion along DNA to escape from its promoter. On the other hand, forms of p53 that are missing the C-terminus activate target genes *in vivo* much more slowly [108] and lack the capacity to resist tumor transformation of cell lines [109]. These results are consistent with recent theoretical work that point to both a negative regulatory effect of excessive non-specific binding through sequestration of transcription factors from their cognate sites and a positive effect of 1D diffusion as part of a mechanism that can greatly reduce the time needed for a transcription factor to reach its promoter [110, 27, 28, 30, 17].

The molecular mechanism underlying 1D translocation of p53 along DNA has heretofore been poorly understood. Two distinct scenarios have been proposed: a *sliding* mode that involves a constant protein-DNA contact, and a *hopping* mechanism that consists of repeated rounds of dissociation and re-association at a nearby location [111]. A high probability of rebinding close to a site of dissociation [112] makes discrimination between the two mechanisms challenging. To distinguish between these two translocation mechanisms, a direct observation of the movement of p53 along DNA is needed. Recent advances in fluorescence imaging have allowed the visualization of individual proteins diffusing along stretched DNA molecules [75, 25]. Here, we report the observation of 1D diffusion of individual p53 proteins along stretched DNA and demonstrate that the protein slides along DNA while maintaining contact with the duplex. We present a quantitative analysis of its

diffusion properties and arrive at a description of the free energy landscape underlying the protein's motion.

4.2 Results

We fluorescently labeled full-length, human p53 and used total internal reflection fluorescence (TIRF) microscopy to visualize its movement along individual λ-phage DNA molecules (Figure 4.2A-D, and *Materials and Methods*). The DNA was tethered at one end to a surface and mechanically stretched by applying a laminar flow of aqueous buffer exerting hydrodynamic drag on the DNA duplex (Figure 4.2) [25].

The fluorescence of the proteins was imaged on a CCD and their positions tracked by determining the Gaussian-fitted center of the single-molecule intensity profiles [78]. Figure 4.2D shows a time series of fluorescence images indicating the movement of an individual p53 along the DNA. Two example trajectories of the movement of individual proteins along the DNA are shown in Figure 4.2E. The mean square displacement (MSD) versus time for the same trajectories is shown in Figure 4.2F.

To estimate the diffusion coefficient of the p53 motion along DNA, we first correct for a drift component in the trajectories due to the hydrodynamic drag exerted by the flow of the buffer on the protein (Figure 4.3). We correct for the effect of drift by subtracting the mean drift over all trajectories (weighted by their durations) from each individual trajectory (see *Materials and Methods*, and Figure 4.4) [113].

The diffusion coefficient for each trajectory then can be calculated by determining the slope of the MSD versus time (*Materials and Methods*). We observe a diffusion coefficient of $(2.60 \pm 2.17) \times 10^6$ bp^2/sec, and and a drift velocity of 262 ± 1144 bp/sec. Figure 4.2G shows a histogram of diffusion coefficients of 162 individual p53 molecules. The

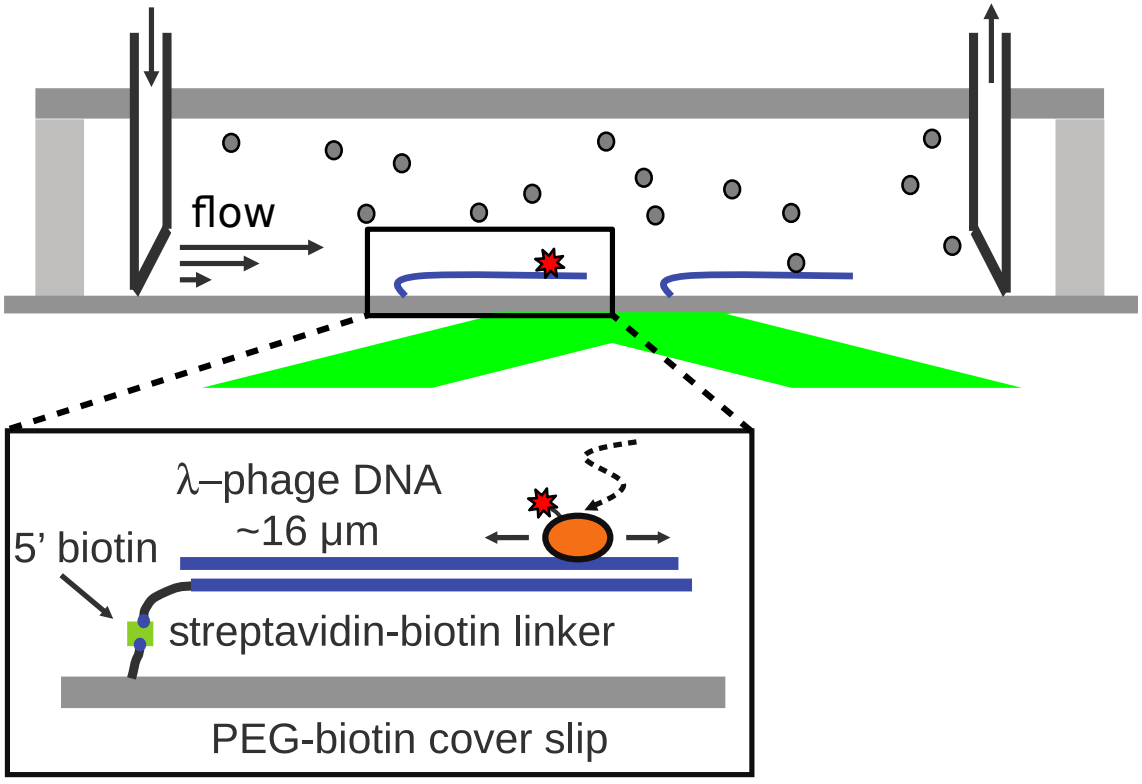


Figure 4.1: Design of the flow-cell. Buffer is pulled into the cell by a syringe pump (not shown), and experiences laminar flow within the cell. The flow stretches λ-DNA molecules which are tethered at one end to the coverslip using a biotin-streptavidin linker. Laser light incident on the coverslip at an angle greater than the critical angle generates an evanescent field, illuminating only those fluorophores that are within 100~200 nm from the surface.

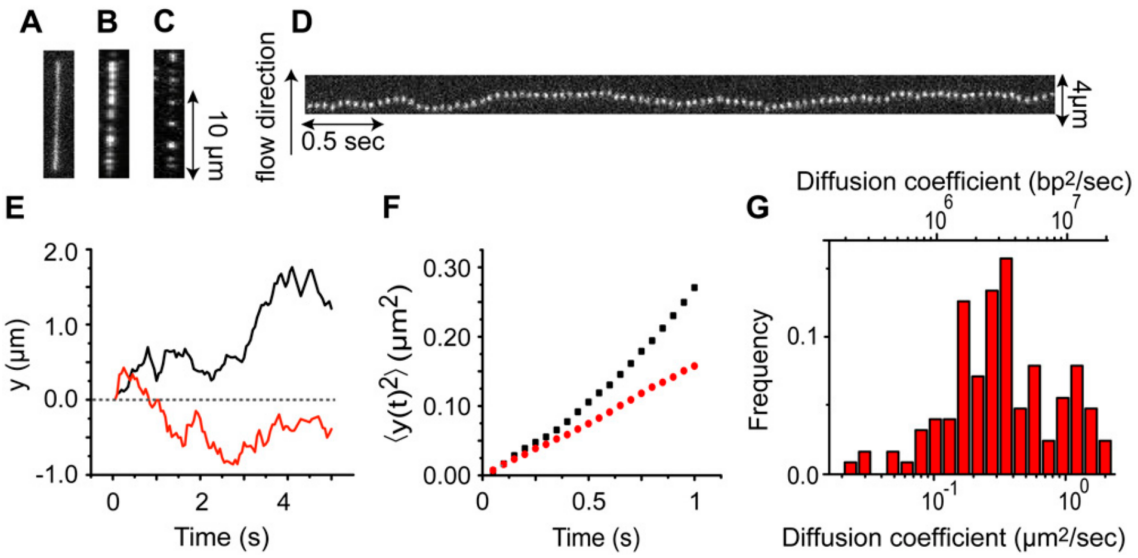


Figure 4.2: Imaging and diffusion coefficients of p53. (A) Stained λ DNA molecule stretched by flow. (B and C) Images of p53 proteins on DNA. Protein concentration is 0.3 nM; the total salt concentration is 75 mM in panel B and 125 mM in panel C. (D) Kymograph of an indi- vidual fluorescently-labeled p53 protein moving on flow-stretched DNA (protein concentration is 5 pM). (E) Diffusion trajectories of two p53 proteins. (F) Mean-square displacement (MSD) versus time of the same two trajectories. (G) Histogram of diffusion coefficient D of 162 individual p53 proteins (125 mM total salt concentration; similar distributions were ob- served with other salt concentra- tions; see Figure 4.5).

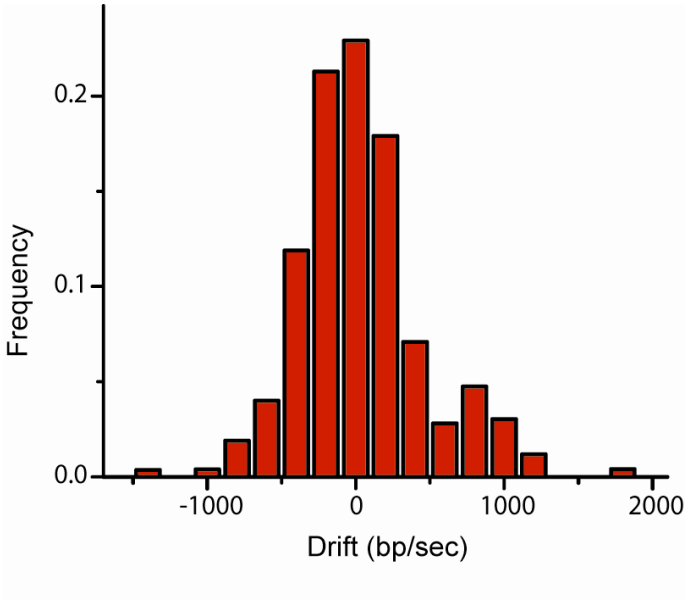


Figure 4.3: Weighted histogram of the drift velocity v of 162 individual p53 proteins. Drift for each p53 protein is calculated by dividing the net displacement of each trajectory by its duration. The weight of each trajectory in the histogram is proportional to its duration. The distribution is biased and the mean of the distribution is at 262 bp/s in the direction of the flow. The depicted results are for 125 mM total salt concentration. Similar distributions were observed for other salt concentration indicating that drift is not dependent on the concentration of salt in our assay.

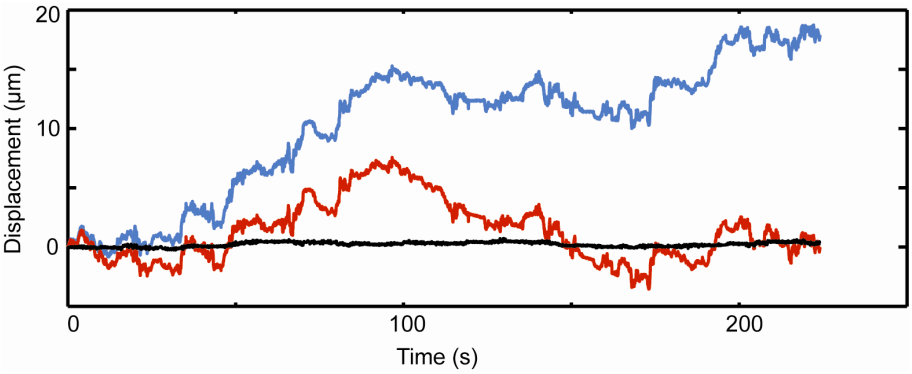


Figure 4.4: Joined trajectories for determining drift. Black line is the displacement in the direction perpendicular to the flow direction. Blue line is the displacement in the direction of the flow and the red line is corrected by reducing the drift effect in the displacement.

large standard deviations do not reflect experimental errors, but describe the width of the measured distributions of diffusion coefficients and drift velocities for many molecules.

Next, we determined whether p53 is moving while maintaining continuous contact with DNA (*i.e.* sliding) or whether it translocates by making small but frequent hops off and back onto the DNA (*i.e.* hopping). Since protein affinity to nonspecific DNA is determined primarily by electrostatic interactions, varying the salt concentration in the experiments can modulate these interactions and allow us to discriminate between the hopping and sliding models [25, 71]. As discussed in section 3.1.2, if a hopping process is responsible for 1D diffusion, a higher salt concentration will lower the nonspecific binding affinity, increasing the fraction of the time the protein spends in solution, and effectively increasing the measured diffusion coefficient. Conversely, a sliding process will result in a diffusion constant that is independent of the salt concentration. Figure 4.5 shows the histograms for diffusion coefficient for hundreds of individual p53 proteins in different salt concentrations. These distributions are summarized in Figure 4.6A (open blue triangles), which shows that the one-dimensional diffusion constants for a range of salt concentrations are indistinguishable, and thereby provides strong evidence against the hopping mechanism and leads us to accept sliding as the principal mechanism of p53's one-dimensional translocation.

In both the sliding and hopping scenarios, the thermodynamic binding affinity of the protein to the DNA is expected to decrease with increasing salt concentration. As a proxy for affinity, we measure the total number of proteins bound to the DNA at various salt concentrations (Figure 4.2B and C; Figure 4.6A, solid red squares) and observe the expected decrease at higher salt concentrations.

Since it had been suggested that the C-terminus of p53 was reponsible for its sliding ability [68], we attempted to examine the diffusive properties of the C-terminus alone. The C-terminal domain is a highly basic, 31-amino acid unstructured domain, containing

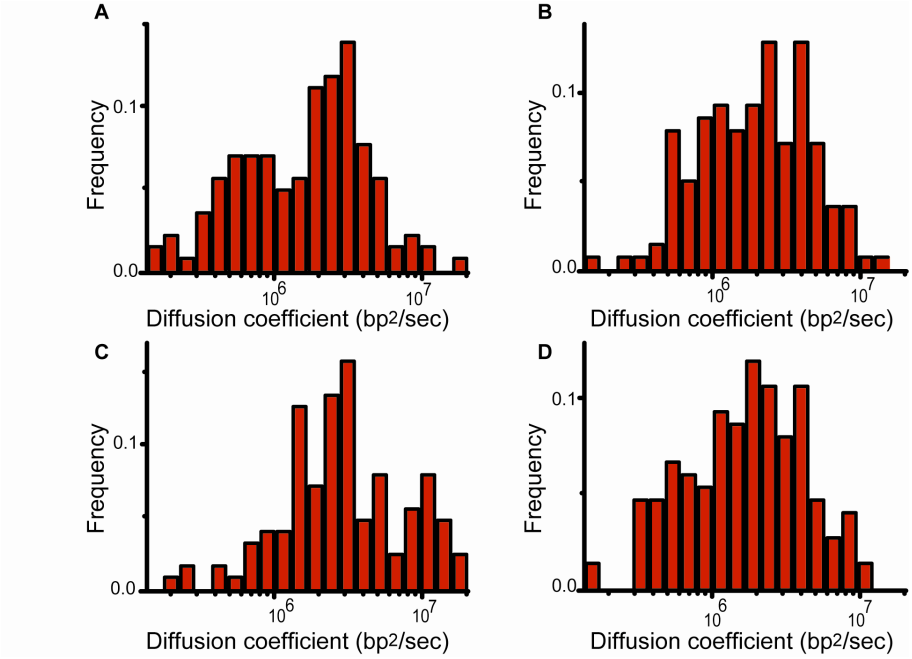


Figure 4.5: The distribution of diffusion coefficients for total salt concentration of (A) 25 mM, (B) 75 mM, (C) 125 mM, and (D) 175 mM. Similar distributions observed for different salt concentration indicate that diffusion coefficient is not dependent on the concentration of salt implying an sliding mechanism for 1D translocation of p53 protein on DNA.

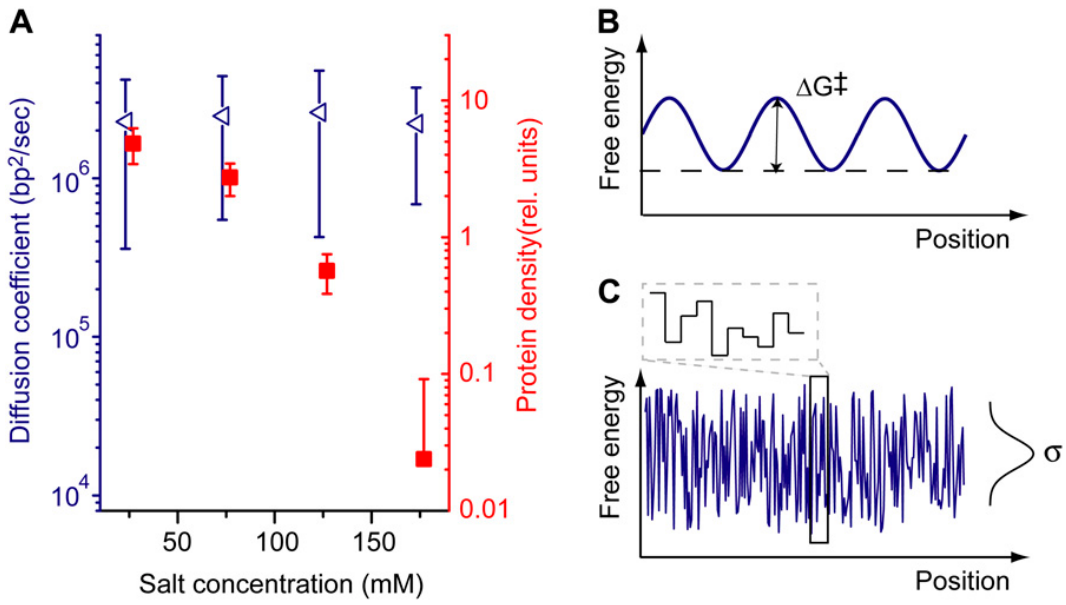


Figure 4.6: (A) Diffusion coefficient D (blue triangles) and protein density on DNA (red squares) as a function of salt concentration. Protein density is measured as the number of observed proteins per kbp of DNA. (B) Iso-energetic model to describe translocation of protein along DNA. For each base-pair, the protein has to overcome an energy barrier of height $\Delta G^{\ddagger}$. (C) Random energy model. Sequence-dependent energies of protein-DNA complex over the length of the DNA follow a Gaussian distribution with variance σ^2.

two arginine and five lysine residues. We fluorescently labeled the C-terminal peptide but were unable to observe it binding at physiological ionic strength (163 mM), which accords with unpublished data from collaborators that at this salt concentration, the affinity of the C-terminal domain is $\lesssim 1000$ times that of the full-length protein. At the same salt concentration, we observe an off-rate of ~1.1/s [114], and so we should expect that the peptide does not bind long enough to be observed with our camera frame rate of 33 Hz.

We found that if we lowered the salt concentration to 13 mM, we indeed observed the C-terminal domain binding to DNA, but the peptide's movement on DNA was indistinguishable from the fluctuations in the DNA polymer itself. Subsequent experiments by our group on a construct that included the tetramerization domain as well as the C-terminal domain would show that this construct indeed slid on DNA and bound enduring enough to record trajectories of [53]. The diffusion coefficient for the tetramerization–C-teriminal construct corresponded to $\sigma = 0.6 k_B T$, suggesting that the C-terminal domain indeed provided p53 with sliding functionality with sufficiently low friction to meet the requirements for efficient sliding under the two-mode model discussed in Part I.

4.3 Discussion

By comparing the experimentally measured diffusion coefficient with the theoretical maximum value for the limiting case of zero protein-DNA friction, we can obtain quantitative information about the free energy landscape of sliding. For a globular protein the size of p53, we estimate the upper limit of the diffusion coefficient to be $D_{lim} = 7.7 \times 10^6$ bp^2/s [71, 115] *Materials and Methods*. Our measured diffusion coefficient of $D_{1D} = (2.60 \pm 2.17) \times 10^6$ bp^2/sec is a factor of 3.6 below this limit. We consider two models that describe this protein-DNA friction.

In the first model, protein-DNA binding energy is constant across all positions (on non-specific DNA), but translocating a distance of one base-pair requires overcoming a free energy barrier of a constant height $\Delta G^{\ddagger}$ (Figure 4.6B). The second model (Figure 4.6C) is a single-landscape random-energy model discussed in Chapter 1, section 1.3.3. The energy of protein-DNA binding varies with the sequence and is normally distributed with variance σ^2, making sliding along DNA a random walk in a random energy landscape. Using the first model, the relation $\langle x^2 \rangle = 2D_{lim}t$, and the assumed step size of 1 bp, we obtain a theoretical upper limit for the stepping rate $k_{lim} = 1.54 \times 10^7$ s^{-1}. From the measured diffusion constant we obtain the stepping rate $k_{exp} = (5.20 \pm 4.34) \times 10^6$ s^{-1}. The Arrhenius relation $k_{exp}/k_{lim} = \exp(-\Delta G^{\ddagger}/k_BT)$, provides a value of $1.78 \pm 1.21 k_BT$ for the activation barrier $\Delta G^{\ddagger}$.

Previous theoretical work demonstrated that the second model yields diffusive behavior with the diffusion coefficient

$$D_{1D} = D_{ideal}(1 + \sigma^2\beta^2/2)^{1/2} \exp(-7\sigma^2\beta^2/4), \tag{4.1}$$

where $\beta = 1/k_BT$ [27] (previously presented as Equation 1.7 in Part I). Using this equation we obtain $\sigma = 0.84 \pm 0.40 k_BT$. Values obtained from the two models are similar and provide a picture of diffusion on a fairly smooth energy landscape, consistent with previous theoretical results that rapid search is possible only with energy barriers $< 2k_BT$ [27]. If the two-mode model discussed in Part I applies to p53, then the time spent in a high-σ recognition (**R**) mode must indeed be small to result in an aggregate σ consistent with a sliding (**S**) landscape.

We have offered the first direct experimental observation of sliding on DNA by p53, and indeed by any eukaryotic transcription factor. One-dimensional sliding is physically necessary for the mechanism of facilitated diffusion, which allows for rapid binding *in vivo* of

transcription factors to their promoters. Further studies will address whether 1D sliding of p53 reported here contributes to facilitated promoter search, a mechanism that is suggested to be available to prokaryotes [71]. Our work opens the way for better understanding of the role of non-specific protein-DNA binding and sliding in negative and positive regulation of gene expression and, broadly, the physical bases of gene regulation. Subsequent to the work presented here, our group examined the role of the various p53 domains and modifications in modulating the kinetics of protein-DNA interactions [53], and the following chapter further tests the applicability of the two-mode model to p53 by investigating sequence-dependence in its sliding kinetics.

4.4 Materials and Methods

4.4.1 DNA preparation and flow stretching

Purified DNA from λ–phage (New England Biolabs) was linearized and biotinylated at one end by annealing a 3' biotin-modified oligo (5'AGGTCGCCGCCC3'-biotin; Integrated DNA Technologies) to the complementary λ-phage 5' overhang. Flow cells (0.1 mm height, 2.0 mm width) with a streptavidin-coated surface were prepared as described previously [116, 117] (Figure 4.2). The streptavidin-coated flow-cell surfaces were blocked by incubation with blocking buffer (Tris 20 mM, EDTA 2 mM, NaCl 50 mM, BSA 0.2 mg/ml, Tween 20 0.005%; pH 7.5) for 20 minutes. Biotin-modified DNA constructs were introduced into the flow cell at a rate of 0.1 mL/min at a concentration of 8 pM for 20 minutes. These conditions resulted in an average density of $\sim$50 surface-tethered DNA molecules per field of view (50×50 μm^2).

The single-molecule imaging experiments were performed in an imaging buffer, containing 20 mM HEPES, 0.5 mM EDTA, 2 mM $MgCl_2$, 0.5 mM DTT, 0.05 mg/mL BSA

(pH 7.9), and varying amounts of KCl. Imaging buffer was drawn into the channel by a syringe pump at a flow rate of 0.1 mL/min, creating shear flow near the coverslip surface [25]. Single-molecule imaging was done with 1–5 pM p53 in imaging buffer; measurements of protein density on individual DNA molecules were done at higher concentrations (100–300 pM).

4.4.2 Protein preparation and labeling

The super-stable mutant of human full-length p53 (fl–p53, residues 1-393) with mutations M133L, V203A, N239Y and N268D [118] was used. Solvent-exposed Cys residues at positions 182, 275 and 227, and the partially buried Cys-124 were all mutated by Ala so that only one exposed Cys (Cys-229) remains. The protein was expressed in *E. coli* and purified as described previously [107, 119]. Cys-229 was labeled with AlexaFluor 488 maleimide from Invitrogen. The labeling was carried out in phosphate buffer (20 mM sodium phosphate, 150 mM NaCl, pH 7.0) with protein concentration of 20–100 μM at 0–4 °C. 10–fold excess AlexaFluor 488 maleimide was added after the disulfide bonds were reduced with 1 mM of tris (2–carboxyethyl) phosphine (TCEP). The labeling progress was followed by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). The reaction was quenched with 2–10 mM β-mercaptoethenol after ~1h and the labeled protein was immediately separated from the free dye on a desalting column. Mass spectrometry analysis of the purified protein ruled out labeling of the protein at stoichiometric excess.

4.4.3 Labeling and troubleshooting of C-terminal peptide

The p53 C-terminal peptide, consisting of amino acids 362-393, with Ser-362 mutated to Cys, was produced using solid-phase peptide synthesis and provided by Dr. Fang Huang.

The N-terminal amine was labeled in a buffer of 100 mM NaCl, 50 mM Na_2HPO_4, and 0.5 mM DDT. A number of labeling conditions were used; the one that produced the labeled peptide used for experiments discussed in *Results* was labeled at a pH of 7.3, at a concentration of 680 μM at 0–4 $^{\circ}C$, with a four-fold excess of AlexaFluor 555 succinimidyl ester. After four hours the reaction was quenched with 0.1% trifluoroacetic acid, and the protein was separated from free dye by dialysis. Spectrophotometry of the products showed the peptide to be labeled with an efficiency of 67%.

No sliding was observed in initial experiments, so we attempted to rule out any artifactual causes of the lack of sliding. In the event that the AlexaFluor 555 dye interfered with sliding, we re-labeled the peptide with tetramethylrhodamine, but observed no change in behavior. To test whether the peptide aggregated, we imaged them in the fluorescence microscope at high laser power to see whether particles photobleached in a stepwise manner. No stepwise pattern was detected , and no difference in photobleaching patterns was observed between the peptides and that had and had not been sonicated immediately before imaging, implying that our peptides were not aggregated and thus aggregation was not responsible for the lack of sliding.

4.4.4 Fluorescence imaging

Fluorescence imaging of the movement of the labeled p53 proteins along DNA was performed by placing the flow cell on top of an inverted microscope (Olympus IX71) and exciting the AlexaFluor 488 label by the 488-nm line from an Ar/Kr laser (Coherent I-70 Spectrum). A high–numerical-aperture microscope objective (Olympus, NA = 1.45) was used to illuminate the sample with total internal reflection. The illuminated area had a diameter of 50 μm at the sample plane. The fluorescence was collected by the same objective and imaged by an EM-CCD camera (Andor iXon), after filtering out scattered laser light. Single-molecule

data was analyzed using custom-written particle-tracking MATLAB code, partially using code obtained from http://physics.georgetown.edu/matlab/.

4.4.5 Particle tracking

The positions of labeled particles were determined by fitting each single-molecule fluorescence image to a two-dimensional Gaussian distribution. The accuracy of position determination is given by

$$\sigma^2 = \left[\frac{s^2}{N} + \frac{a^2/12}{N} + \frac{8\pi s^4 b^2}{a^2 N^2}\right] \tag{4.2}$$

where N is the number of photons collected [78]. Typical signals from individual AlexaFluor 488 labels corresponded to 125±56 photons per 50-ms integration. Using the standard deviation of the microscope point-spread function s (140 nm for our microscope), the pixel size a (166 nm), and the standard deviation of the background level b (20 photons), we calculate the standard error of position determination to be σ= 10–20 nm.

4.4.6 Determination of drift rates and diffusion coefficients

We evaluate the presence of any directional bias in protein motion by measuring the net displacement of a protein divided by the duration of its trajectory. In the absence of any drift, the net displacement of a population of molecules undergoing normal diffusion will form a normal distribution around zero. In our experiment, however, we observe a small bias of the proteins' motion in the direction of the flow. The flow-induced drift distances are about a factor of 5 smaller than the diffusional distances at experimental timescales and are likely to have a minimal impact on the analysis of the diffusion properties of the protein. Nonetheless, we evaluate the effect of drift and diffusion as separate contributions by subtracting the mean drift over all trajectories at a particular biochemical condition from that condition's individual trajectories and calculating the diffusion coefficient from

the drift-corrected trajectories [120, 113]. This method assumes that drift and diffusion are independent and that their contributions to the displacement of the protein is additive.

The mean drift is the mean of each trajectory's total displacement over its duration, weighted by the duration of the trajectory. This weighted mean is equivalent to the total drift divided by the total duration of all trajectories, as if they were concatenated into a single long trajectory (Figure 4.4, blue line). The standard deviation of the drift for a given biochemical condition is likewise weighted according to the durations of the trajectories. As shown in Figure 4.3, for 162 sliding proteins at 125 mM total salt concentration, the distribution of the drift component of the trajectories is shifted 262 ± 1144 bp/sec from zero in the direction of the flow.

Having experimentally obtained trajectories of multiple particles, we determine the diffusion coefficient D of particle of N frames by plotting the mean square displacement (MSD) of the particles as a function of time windows $n\Delta t$, and fitting the resulting data to a straight line, whose slope equals $2D$.

$$\mathrm{MSD}(N,n) = \frac{\sum_{i=1}^{N-n} (y_{i+n} - y_i)^2}{N-n} = 2Dn\Delta t. \tag{4.3}$$

We measure the diffusion coefficient by the slope of the fit to the data corresponding to $n = 3 - 10$ (0.15–0.5 seconds). The upper bound is limited by the typical length of the trajectories, whereas the lower bound is chosen to exclude the effect of short-lived DNA fluctuations. The DNA fluctuations appear on a timescale of less than 0.1 second as determined by tracking particles that appear to not slide on the DNA. The MSD plot of such particles is linear only on timescales less than 0.1 second and shows bounded diffusion for longer timescales. Therefore, by excluding small time windows, we are avoiding the fluctuations of DNA to appear in the calculation of the diffusion coefficient. Figure 4.7 shows the MSD data for different proteins. The majority of the proteins display linear MSD versus

time plots, indicating normal Brownian diffusion along DNA.

Out of 484 initial trajectories for 125 mM total salt concentration, 327 have trajectories longer than 1.5 seconds, with the remainder being too short to result in reliable diffusion coefficients. Out of these 327, the 235 trajectories with total distance traveled longer than 500 nm are chosen to avoid particles non-specifically bound to the glass surface of the flow cell, as well as particles on DNA that are not sliding. The MSD curves for the majority of these proteins are linear. To avoid fitting curves corresponding to nonlinear MSD vs. $n\Delta t$ curves, we only take into consideration the 162 molecules for which the Pearson correlation coefficient between MSD and $n\Delta t$ is greater than 0.9. Trajectories with non-linear MSD curves are likely to be non-sliding proteins on DNA with high-amplitude fluctuations. Diffusion and drift coefficients were determined from these 162 final trajectories in 125 mM total salt concentration. Similar proportions of proteins were selected in each of the above steps for other salt concentrations, and the number of final analyzed trajectories was similar across different biochemical conditions. Also, the overall shape of the distribution of diffusion coefficient is similar for different salt concentrations (Figure 4.5).

4.4.7 Calculation of activation-barrier heights in sliding

The Stokes-Einstein relation (Equation 3.1 in section 3.1.2, reprinted here for convenience) gives the 1D diffusion coefficient for a spherical object diffusing by purely translational movement.

$$D_{1D,lim} = \frac{k_B T}{6\pi\eta a} \tag{4.4}$$

where η is the solvent viscosity (8.9×10^{-4} Pa·s for water at 25 °C), a is the radius of the diffusing p53 protein (3.9 nm [121]), k_B is the Boltzman constant and T is the temperature. For p53, this calculation results in a one-dimensional diffusion coefficient of 6.3×10^8 bp^2/s.

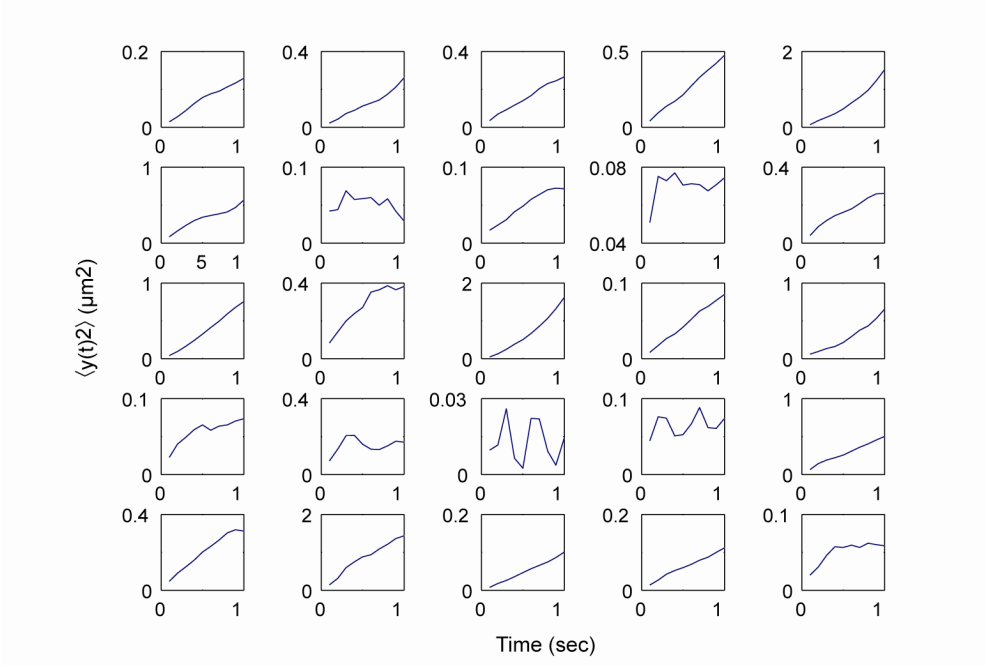


Figure 4.7: Mean square displacement vs. $n\Delta t$ ($\Delta t = 0.05$ s) for different p53 particles at a total salt concentration of 125 mM. As can be seen from the plot, most of the trajectories show linear dependence of MSD on time, indicating normal Brownian diffusion along DNA. Out of 235 trajectories, 162 were selected based on the selection criteria described in the text.

When Equation 4.4 is modified to require the protein to track the DNA helix (Equation 3.2), an upper value of $D_{1D,lim}$ of 7.7×10^6 bp^2/sec is obtained. Assuming a step size of 1 base pair, the upper limit of the stepping rate can be calculated as [25, 122]

$$\frac{2D_{lim}}{< x >^2} = k_{\text{lim}} \tag{4.5}$$

which results in an upper limit for the stepping rate k_{lim} of 1.54×10^7 steps/sec. The measured diffusion constant D_{exp} has a value of $(2.60 \pm 2.17) \times 10^6$ bp^2/sec, corresponding to a stepping rate of $(5.20 \pm 4.34) \times 10^6$ steps/sec. From the Arrhenius relation

$$\frac{k_{\text{exp}}}{k_{\text{lim}}} = \exp\left(\frac{-\Delta G^{\ddagger}}{k_B T}\right) \tag{4.6}$$

We calculate a value for the height of the activation barrier, $\Delta G^{\ddagger}$, of 1.78 ± 1.21 $k_B T$ for the protein-DNA constant–energy-barrier model.

4.4.8 Stokes drag force

In order to calculate the Stokes drag force exerted on a protein bound to the DNA, we need an estimate for the velocity of the buffer flow at the position of the DNA-bound protein. In our flow-stretching, the buffer solution was drawn into the channel by a syringe pump with a flow rate of 0.1 mL/min creating shear flow near the cover slip surface. The flow channel is 100 μm in height and 2 mm in width, resulting in an average velocity of the buffer of 0.83 cm/sec. The flow velocity, however, is not constant throughout the channel, but is zero at the boundaries, yielding a parabolic flow profile [122]. The mean distance of the DNA from the coverslip surface is 0.2 μm [25, 122, 123]. With a channel height h, the flow velocity v_y at a distance y from the surface can be expressed as:

$$v_{avg} = \frac{2}{3} v_{max}, \quad \text{and} \quad v_y = v_{max}\left[\frac{hy - y^2}{h^2/4}\right] = \frac{3}{2} v_{avg}\left[\frac{hy - y^2}{h^2/4}\right] \tag{4.7}$$

The average velocity of the flow at the center of the DNA, 0.2 μm above the surface, can be estimated as 100 μm/sec. The Stokes drag force exerted on an object close to a surface is given by

$$F = 6\pi\eta r v \left(1 + \frac{9r}{16y}\right) \tag{4.8}$$

with η denoting the viscosity, r the radius and y its distance from the surface [124]. The force exerted on a single p53 bound to the DNA at a distance of 0.2 μm from the surface is calculated to be $\sim$6.6 fN and is responsible for the bias in protein translocation in the direction of the flow.

4.4.9 Measuring the protein density on DNA

To measure the dependence of the binding affinity of p53 for nonspecific DNA, we counted the number of DNA-bound proteins per λ-DNA molecule and divide by the DNA length to obtain a protein density (Figure 4.2B and C). The high sensitivity of binding affinity to salt concentration made it difficult, however, to choose a protein concentration that allows for an unambiguous determination of the number of molecules at the various salt concentrations used. Instead, we measured the number of detected photons per unit length of λ–DNA as a proxy for the number of proteins bound. The single-molecule sliding experiments provided an average intensity per p53 protein of 125$\pm$ 56 photons/sec, a value that was used to convert intensity per unit length of DNA into number of proteins per unit length of DNA.

4.5 Acknowledgements

Most of the text and nearly all of figure material in this chapter is taken from

> Tafvizi A, Huang F, Leith JS, Fersht AR, Mirny LA, van Oijen AM (2008) Tumor Suppressor p53 Slides on DNA with Low Friction and High Stability. *Biophys. J.* **95**:L01-L03 (citation [43]).

The paper was written largely by A.T., with contributions from J.L., as well as from A.v.O. and L.M. Other than the labeling of the C-terminal domain, the experiments were carried out by A.T., and the protein provided by F.H. A.T. and J.L. performed the data analysis, though the lion's share of the credit redounds to A.T. The authors thank Dr. Satoshi Habuchi and Paul Blainey for technical advice and helpful discussions. L.M. acknowledges support of the NIH-funded National Center for Biomedical Informatics i2b2 and thanks the organizers of the meeting on Protein Assembly, Dynamics and Function at IHES which was essential for establishing this collaboration.

Chapter 5

Sequence-dependent sliding kinetics of p53 on DNA

The work discussed in the previous chapter established that tumor suppressor p53's sliding behavior on DNA satisfies a number of conditions of the two-mode model for protein-DNA search and recognition, as laid out in Part I. The model requires that the protein's 1D translocation proceed at least partially by sliding, that is, maintaining contact with the DNA, rather than hopping, in which the protein visits a single position on DNA and then returns to solution. It also requires that the protein slide with $\lesssim 1 - 2k_BT$ of ruggedness. The previous study demonstrated that p53 indeed slides rather than hops, and that it does so with $< 1k_BT$.

This chapter discusses subsequent work that more definitively demonstrates the applicability of the two-mode model to p53. For sliding to be an effective means of accelerating target localization, the protein must sample the sequence of the DNA on which it slides. This sampling may take the form of visits to the recognition (**R**) mode, or sliding (**S**) landscape with ruggedness correlated to that of the **R** landscape, or both. In the first

case, the time spent in the **R** mode will be dependent on the **R**-mode affinity to the particular sites sampled, and in the second case, the time spent in potential wells in the **S** landscape will depend on the frequency and depth of traps or target sites in the **R** landscape. In either case, the sliding kinetics of the protein can be expected to be a function of the DNA sequence. Using mostly similar experimental techniques as described in Chapter 4 and novel data analytical techniques, we determined that p53's diffusion coefficient, D, is indeed sequence-dependent, and that the variance in D among regions of a long (~50,000 base pairs) DNA molecule correlates with predicted variance based on the two-mode model.

5.1 Introduction

As discussed previously in this Part, p53 is a critical transcription factor in preventing tumorigenesis. In addition to its clinical importance, it is noteworthy for being the first eukaryotic transcription factor (TF) directly observed to undergo one-dimensional (1D) diffusion on DNA[43], described in Chapter 4. This 1D sliding has long been hypothesized to facilitate the diffusion of passively-transported site-specific DNA-binding proteins (DBPs) to their targets on DNA [125, 126] (see Chapter 1), and was more recently demonstrated in bulk biochemical experiments to play a role in p53's activation of target genes [68, 127]. Several theoretical and experimental studies[19, 21, 61, 42] have shown that despite a vast excess of accessible DNA (10^7–10^9 bp) to which DBPs have non-specific affinity, their search process can be efficient if they alternate rounds of 1D sliding while bound non-specifically with rounds of 3-dimensional (3D) diffusion in solution between different sections of genomic DNA. Until recently such studies have been limited to bacterial systems [42], and it remained unclear whether the same mechanism was at play in eukaryotes where DNA is packed by nucleosomes limiting space for sliding. Recently Larson *et al.*[52] have demon-

strated that yeast DBPs search for their sites by a 1D/3D mechanism. By demonstrating the ability of p53 not only to slide but also to "read" DNA while sliding, our study provides strong support for 1D/3D mechanism in high eukaryotes.

For its sliding to be functional in promoting efficient search, a DBP such as a transcription factor must be able to read the DNA sequence while sliding along it. This functionality implies that the binding energy at each DNA position depends on the sequence. The magnitude of this sequence dependence can be captured by the standard deviation of the energies comprising the landscape, σ. Smaller values of σ correspond to less-rugged landscapes and thus the ability of the protein to sample sites on DNA more rapidly. Prior theoretical work [39] demonstrated that $\sigma \lesssim 1.5 k_B T$ is required for sliding to be effective in facilitating the search, and at the same time, pointed out that stability of the protein-DNA complex requires $\sigma \gtrsim 5 k_B T$. These mutually unsatisfiable requirements lead to an apparent "speed-stability paradox"[39, 37], which had been qualitatively anticipated [128].

We proposed a multi-mode model of protein-DNA interaction to resolve this paradox[39, 37] (see Chapter 1). To review, in the model's simplest form, a site-specific DBP exhibits two modes of binding between which it stochastically switches: a *search*, or **S**, mode, and a *recognition*, or **R**, mode, with respective **S** and **R** energy landscapes (illustrated in Part I as Figure 1.3A,B; reprinted here for convenience as Figure 5.1A,B). In the **S** mode, the protein binds with sufficiently small sequence-dependence ($\lesssim 1.5 k_B T$) to slide efficiently. In the **R** mode, the protein binds highly-specifically and sliding is negligible. The modes differ physically presumably in their conformations, the **S** mode having chiefly backbone interactions, for example, while interactions with nucleobases present in the **R** mode. For the two-mode system to be effective in providing both rapid sliding and efficient recognition of the cognate site, the **S** mode must have significantly lower average energy and thus be favored at nearly all binding positions to avoid unproductive visits to **R** mode at the

vast majority of sites that do not resemble the cognate site [37]. Transition into **R** mode at the sites that have low energy, and thus resemble the cognate site, slows down sliding. The central idea of this study is that such slow-down on near-cognate site can be detected experimentally.

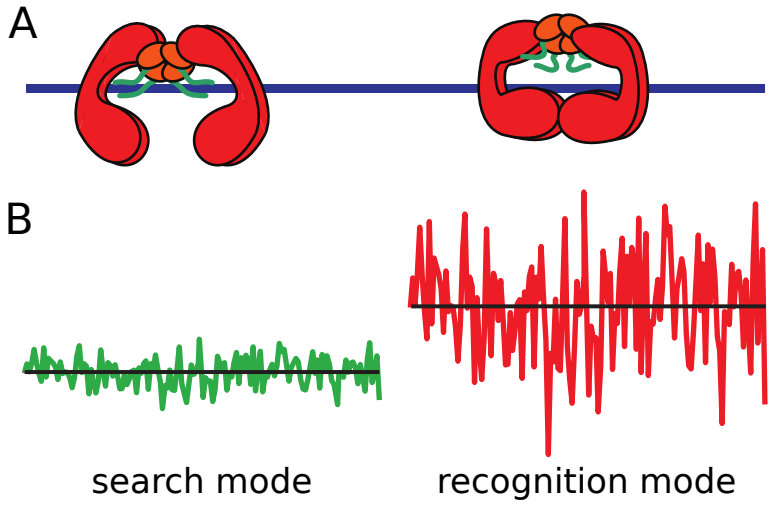


Figure 5.1: Cartoons and energy landscapes of p53 on DNA in search (**S**) and recognition (**R**) modes. **(A)**: Cartoon model for p53, based on EM data[47], indicates the domains responsible for the modalities: green C-terminal domain for the **S** mode; red core domain for the **R** mode. Tetramerization domain shown in orange. **(B)**: Energy landscapes. In **S** mode, the protein interacts chiefly with the DNA backbone and experiences a smooth landscape. In **R** mode, it interacts with the nucleobases, yielding a highly sequence-dependent landscape.

Structural evidence for the two-mode model has been discussed in section 1.3.4. For p53 in particular, electron microscopy measurements on p53 bound to an oligonucleotide with a cognate sequence flanked by non-specific DNA has identified multiple binding conformations [47]. Additionally, studies conducted in by our groups on p53 truncation mutants have shown that distinct domains—the C-terminal domain and the core domain—are responsible respectively for p53's sliding and recognition functionalities[53] (Figure 5.1A). The C-terminal domain indeed is estimated to experience an energy landscape with $\sigma \approx 0.6 k_B T$, satisfying the requirements for an efficient search landscape, while the specifically-binding core domain cannot slide on its own.

Here we report measurements using single-molecule fluorescence microscopy of p53's sequence-dependent diffusivity. We observe that the transcription factor's sliding kinetics on λ-phage DNA in the absence of known cognate sites vary by a factor of approximately 1.6 among different regions of DNA. Using a model with both **R** and **S** modes and a model with only a single mode, we construct predicted effective energy landscapes for the protein on DNA and demonstrate that the two-mode model but not the one-mode model accounts for the observed variation in diffusion coefficient among regions of the DNA. We further provide evidence that the two identical homodimers making up biologically active tetrameric p53 can bind DNA in dissimilar modes, i.e., "hemi-specifically". Such binding has been observed in other DBPs binding to oligonucleotides containing "half-sites", that is, sequences that amount to one half of its recognition sequence[50]. Our analysis of p53's sequence-dependent sliding kinetics reveals that the hemi-specific binding is a general feature of p53's interaction with DNA, with between a fifth and a quarter of the sequence dependence in the protein's sliding kinetics owing to hemi-specific interactions with half-sites.

5.2 Results

To assess whether p53 diffusivity varies depending on its position on λ-phage DNA, we recorded trajectories (Figure 5.2) of fluorescently-labeled single p53 molecules on DNA that was tethered to the surface of a flow cell and stretched by shear flow, using total internal reflection fluorescence microscopy (previously described in Chapters 3 and 4; see Figure 4.2). We mapped trajectories to the contour of DNA (Figure 5.3A and in *Materials and Methods*, Figure 5.8) and determined maximum likelihood estimates of diffusion coefficients, D, of p53 particles, while taking into account drift from buffer flow

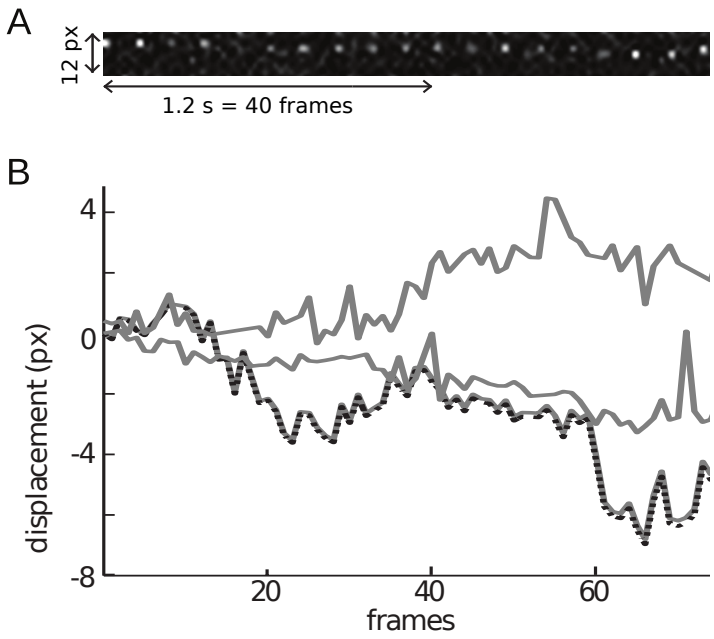


Figure 5.2: Measurements of p53 sliding on DNA, initial data analysis. **(A)**: Kymogram of a single p53 protein diffusing on DNA. Flow direction is up; every fourth frame is shown, giving an apparent frame rate of 120ms. **(B)**: Trajectories of three particles (gray). The dotted black trace represents the bottommost trajectory corrected for drift. The bottommost trajectory corresponds to the kymogram in (A).

and position-dependent DNA fluctuations measured using quantum dots (*Materials and Methods* 5.4.2). This approach shows that DNA fluctuations cannot account for observed particle diffusivity: the square of the central 95% of the range of the particles exceeds the amplitude of the square displacement from DNA fluctuations in the limit of long time windows, Δt, by 1–3 orders of magnitude (Figure 5.3B). The diffusion coefficient for each p53 trajectory, along with its range covered on the DNA, is shown in Figure 5.3C.

We observed that different regions of the λ-phage DNA correspond to different diffusion coefficients. We determined an aggregate experimental diffusion coefficient, D_{expt}, for each segment by assigning every midpoint of each particle trajectory displacement to a position on the DNA, binning the contour of the DNA into ~3-kb segments, and calculating the mean diffusion coefficient within each segment (Figure 5.3A,C,D). Error bars in Figure

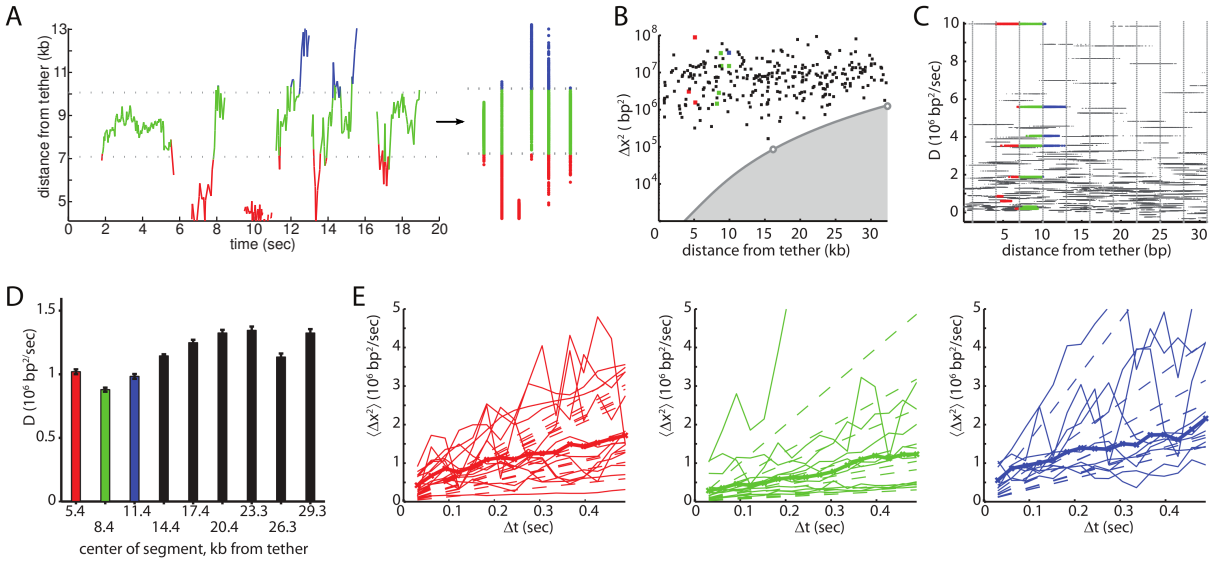


Figure 5.3: Data analysis: Diffusion coefficients of p53 on λ-phage DNA. **(A)**: Trajectories of selected particles in three representative segments. The trajectories have been spread out along the horizontal axis for clarity. Portions of trajectories are colored according to the segement in which they lie: red, green, and blue for segments 1, 2, and 3. For each particle, the positions and assigned segments of each displacement are shown to the right. **(B)**: Square markers: Squared range of the central 95% of each trajectory, plotted over the trajectory's midpoint. Colored markers correspond to particles shown in (A). Gray circles: mean squared displacement at long (>100ms) time windows of quantum dots fixed at 1/3 and 2/3 the contour distance from the tether point. Shaded region is the mean squared displacement of the DNA at long time windows. **(C)**: Horizontal lines consist of dots plotted on the horizontal axis at the midpoint of each displacement within a trajectory, and on the vertical axis at the D of their respective particle. Colored dots correspond to colored dots in (A). **(D)**: Estimated D for each segment. Colored bars correspond to coloring scheme for (A)–(C). Uncertainties were determined by bootstrapping: the displacements for each segment were resampled 1000 times and the resulting diffusion coefficients calculated. Error bars represent a standard deviation in the resampled diffusion coefficients above and below the estimated D. **(E)**: Thin solid traces are MSD/Δt for particles whose median position lies within segments 1 (red), 2 (green), or 3 (blue). For clarity, only every third particle is shown. Dashed lines are $2D$ for that particle as determined in *Materials and Methods*. Thick solid traces are the weighted mean MSD/Δt for the particles shown.

5.3D are errors of estimated D_{expt} calculated as standard deviation of D from 1000 bootstrap resamples of the displacements. We found that segments' aggregate D spanned a factor of approximately 1.6, with 10 of the 36 pairs of segments differing in D significantly at $\alpha = .05$ and 5 of the pairs at $\alpha = .01$ (Figure 5.4). Plots of the mean-square displacement (MSD) as a function of time window Δt for particles in selected segments are shown in Figure 5.3E.

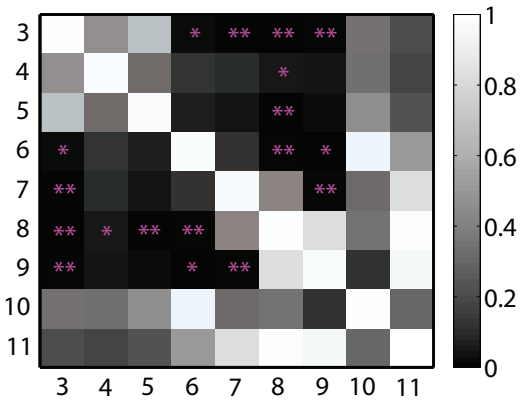


Figure 5.4: P-values of the ratio of D between pairs of segments. p $<$ .05 denoted by *; p $<$ .01 denoted by **. P-values were computed by randomizing the assignment of each particle's diffusion coefficient D as determined in *Materials and Methods* section 5.4.2 to the midpoints of the particle's displacements, and then computing D for each segment. This procedure was repeated 1000 times and for each pair of segments i and j, the frequency of observing the absolute log-ratio, $|\log(D_i/D_j)|$, greater than the absolute log-ratio between the respective Ds from the unrandomized data set was determined.

Next we tested whether this variation in D_{expt} could be explained by sequence-specific binding in the **R** mode. Since λ-phage DNA contains many sites that resemble half- and full-sites of p53, we expected p53 could bind these sites thus slow down sliding. To this end, we developed a model of an effective two-mode landscape experienced by tetrameric p53 on DNA. Experimental studies have demonstrated that p53, a dimer of dimers with a response element (RE) of 20 bp, binds with one of its dimers to a 10-bp half-site with greater affinity than to random DNA [91, 88]. Accordingly, we posited that each dimer could bind a position on DNA in the **R** mode with an energy depending on the sequence,

$E_R(x)$, or in the **S** mode with constant energy E_S, while the other dimer could bind in a similar (fully specific or fully non-specific) or dissimilar mode (hemi-specific) (*Materials and Methods* Section 5.4.3; Figure 5.5). A cooperativity term ϵ is included to account for additional binding energy when both dimers bind in **R** mode.

We calculated a sequence-specific binding landscape $E_R(x)$ using a position weight matrix (PWM) for a single dimer, based on known p53 REs [97] (Figure 5.6A). Six other lists of p53 REs produced very similar sequence logos to the one used for our study (*Appendix* 5.A6). Then, an effective binding energy for the tetramer, $U(x)$, was computed over all positions according to the two-mode model (*Materials and Methods*, Equations 5.5–5.8), as well as according to a single-mode model, giving rise to respective energy landscapes (Figures 5.6E and 5.6C). The calculation of $U(x)$ allows variable spacing between the 10-bp sequences bound by the two dimers. We identified sites of λ DNA that scored as well as some the weaker known p53 REs, but are not known to be *in vivo* targets of p53 (Figure 5.6B). The PWM was scaled to fit experimentally measured dissociation constants of p53 and oligonucleotides containing full-sites, half-sites, and random sequences [88]. The difference between E_S and the average of $E_R(x)$ was set so that the free energy difference between specific and non-specific binding for typical eukaryotic TFs [129] would match the difference for our landscape.

From the computed landscape, we predicted each segment's reduction in diffusivity relative to a featureless **S** landscape, D/D_0 (*Materials and Methods* and *Appendix*). Areas with more/deep energy wells were found, as expected, to correspond to reduced diffusivity of particles in these areas (Figure 5.7A). We demonstrate that D/D_0 is ratio of the time t_s a protein spends sliding in **S** mode to the total time the protein spend on the landscape, t_{total}:

$$\frac{D}{D_0} = \left\langle \frac{t_s}{t_{total}} \right\rangle = \frac{n \exp(-2E_s)}{\sum_x^n \exp(-U(x))} \tag{5.1}$$

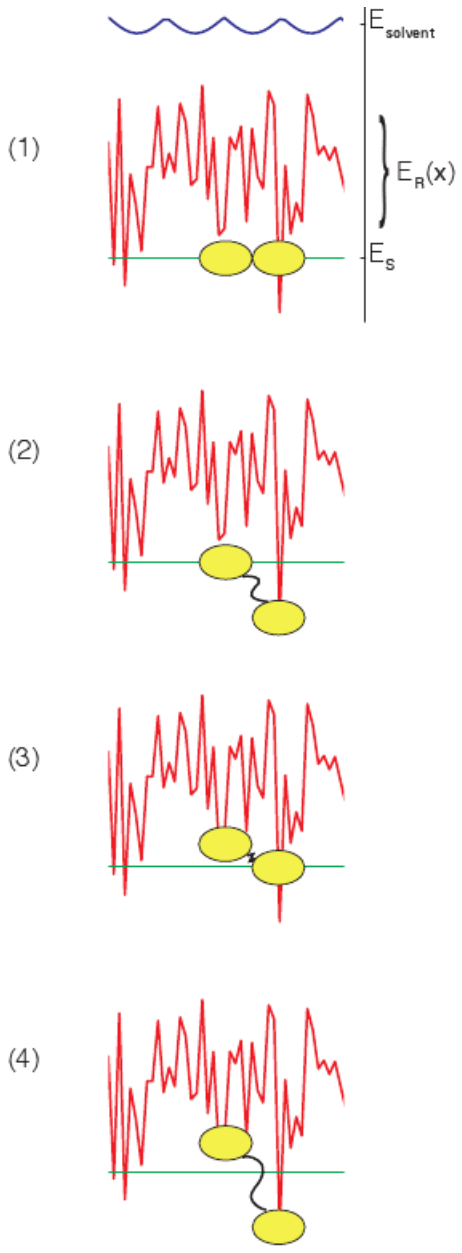


Figure 5.5: Four modes of binding: (1) fully non-specific; (2) first dimer non-specific, second dimer specific; (3) first dimer specific, second dimer non-specific; (4) fully specific. The energy at a position x in the golf-course landscape is equal to the negative logarithm of the sum of the statistical weights of these four modes.

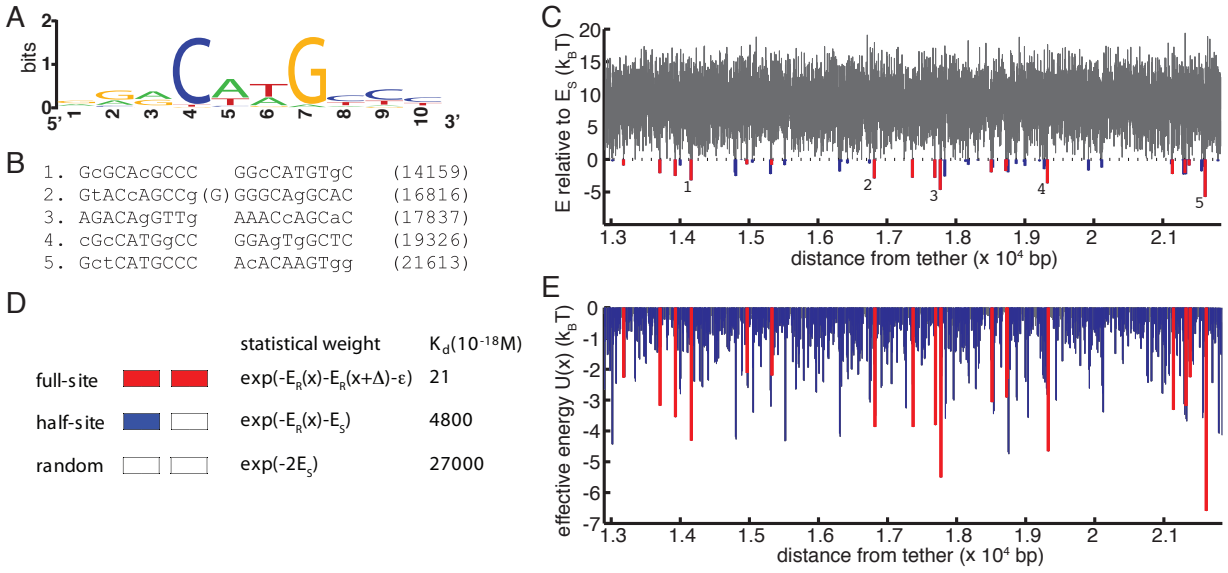


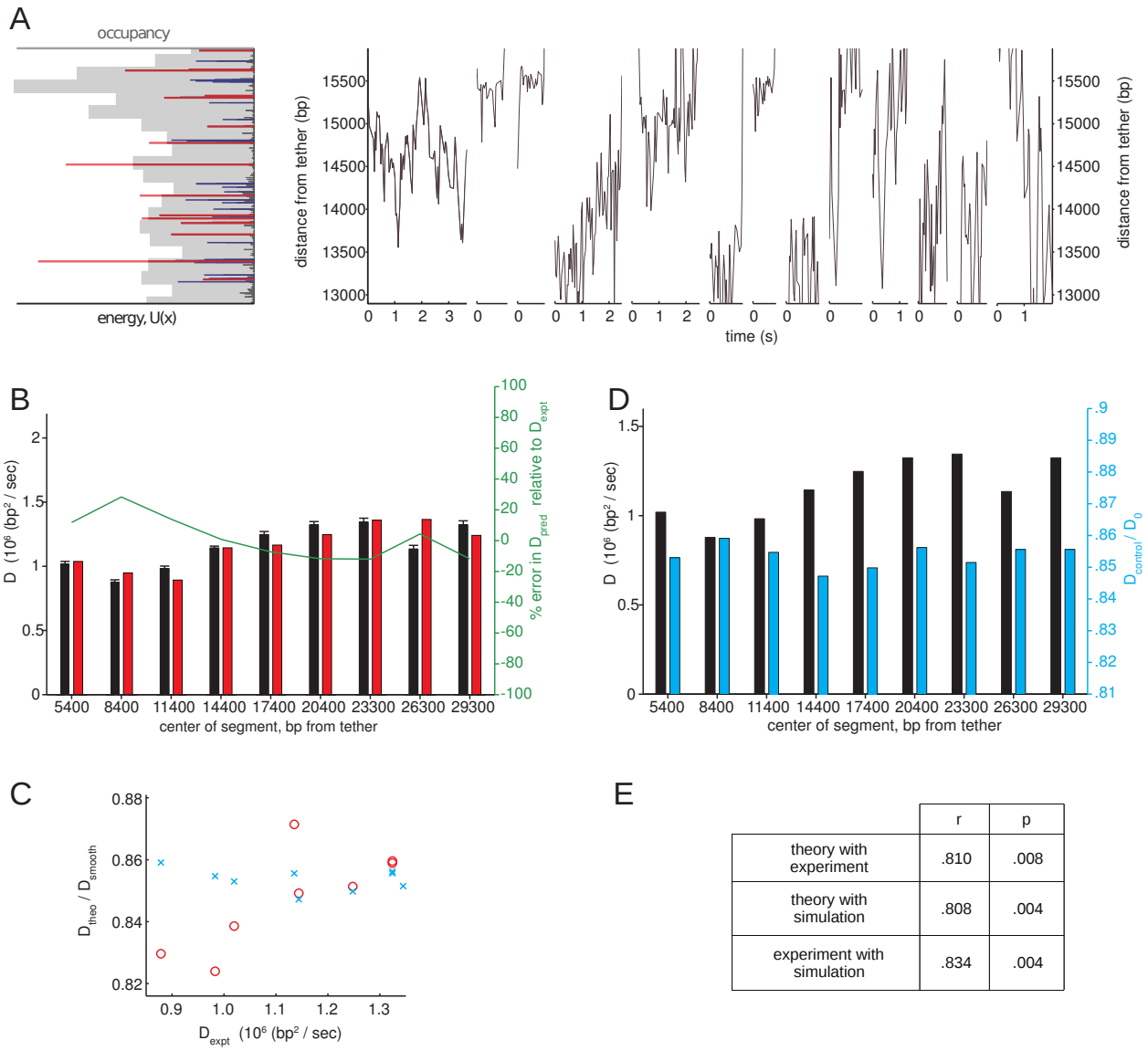
Figure 5.6: Theory: scoring the λ genome and predicted landscapes. **(A)**: Half-site sequence logo for p53. **(B)**: Sequences and positions in bp from the tether of full-sites found in segments 4–6 of λ DNA shown in (C) and (E). Lowercase letters indicate nucleotides that do not match the consensus sequence of RRRCWWGYYY. **(C)**: Predicted one-mode landscape in segments 4–6. Full-sites are colored red; half-sites colored blue. **(D)**: Table describing key elements of two-mode model. The statistical weights of fully-specific, hemi-specific, and non-specific binding at a position x in making up $U(x)$ (Equation 5.5) are indicated. For the great majority of positions on DNA that lack a half-site, the greatest of these is the term representing fully-non-specific binding. For positions that include a half-site, it is the term representing hemi-specific binding. For full-sites, it is the term representing fully-specific binding. K_d's are of p53 to representative examples of full-sites, half-sites, and non-sites[88]. **(E)**: Predicted two-mode landscape. Most positions are dominated by non-specific binding. The possibilty of hemi-specific binding makes half-site binding relatively more important for the two-mode model than for the one-mode model.

This result is based on the assumption that traps are isolated or that the protein does not slide in the **R** mode.

We compared D_{expt} with D/D_0 over the segments, and found the experimental and predicted diffusion coefficients to correlate strongly ($r = .81$, $p = .008$) (Figure 5.7B, black and red bars). To determine the significance of this correlation, r_{expt}, we implemented a permutation test. We constructed 500 scrambled landscapes of $E_R(x)$, computed $U(x)$ and D/D_0 for each of them, and determined the resulting r_{ctl} between the their predicted D/D_0 and the experimental D_{expt} over the segments (an example in Figure 5.7C,D). The correlation between predicted D/D_0 and D_{expt} exceeded that between control D/D_0 and D_{expt} for all but 4 of the 500 control landscapes ($p = .008$). The observed strong and significant correlation demonstrates that a two-mode sequence-specific landscape can explain the observed positional variability of p53's diffusion coefficient.

We also compared the ruggedness of effective landscape $U(x)$, formed from an **R** and an **S** landscape, with earlier experiments and with theoretical requirements. Satisfyingly, the global ruggedness σ of the two-mode landscape is $0.51k_BT$, which lies below the theoretical upper limit for efficient search, $\sim 1.5k_BT$, and falls within the uncertainty for the aggregate σ, $0.84 \pm 0.40k_BT$ obtained for p53 earlier [43]. In contrast, the landscape without a non-specific binding mode has $\sigma = 3.5k_BT$, which is too great on theoretical grounds for efficient sliding and moreover is incompatible with observed diffusion coefficients. Furthermore, the diffusivity D/D_0 computed for one-mode landscape shows no significant correlation with D_{expt} ($r = .51, p = .11$). This allows us to rule out the one-mode model.

Experimental observations of dimeric p53 binding to 10-bp half-sites[91] prompted us to explore the role of hemi-specific binding of tetrameric p53. The two-mode model discussed thus far does not require that the two dimers making up the full tetramer bind in



	r	p
theory with experiment	.810	.008
theory with simulation	.808	.004
experiment with simulation	.834	.004

Figure 5.7: Comparison of theory, simulations, and experiment. **(A)**: Center: trajectories within segment 4, ordered by increasing estimated D. Left: Red and blue bars denote predicted potential wells, with height of bars proportional to predicted effective energy, $U(x)$. Full-sites are colored red; half-sites are colored blue. Gray bars are a histogram of particle occupancy within the segment, with bin widths equal to one twentieth the segment width. **(B)**: Estimated D for experimental (black bars) and predicted D/D_0 (red bars) for the predicted landscape, over segments along λ DNA. D/D_0 is scaled to match D_{expt}'s mean and coefficient of variation. Green trace is the percent error in predicted D/D_0 normalized by the mean D/D_0, relative to D_{expt} normalized by the mean D_{expt}. Uncertainty in estimates for D from simulations were determined by the same bootstrapping technique used for experimental D described in Figure 5.3. **(C)**: Scatter plot of D_{expt} versus D/D_0 for all segments. Red circles correspond to values for the predicted landscape based on the two-mode model; cyan x's correspond to values for the control landscape whose correlation with D_{expt} was the median from among the 500 control landscapes. **(D)**: Black bars are identical to those in (B). Cyan bars correspond to D/D_0 of the control landscape that produced the cyan x's in (C). **(E)**: Correlation coefficients and p-values.

the same mode. This gives the protein enhanced affinity for half-sites even when the half-site is flanked by sequences that would be unfavorable to bind in the **R** mode (Figure 5.6D). We found that the segments' experimental diffusion coefficients correlated more weakly with predicted D/D_0 when we eliminated the possibility of hemi-specific binding—tantamount to removing the two middle terms of Equation 5.5—than they did for the full-model D/D_0 ($r_{\text{no hemi}} = .72$ versus $r = .81$). The fraction of the sequence-specificity of p53's diffusion coefficient that owes to full-sites is thus approximately $\frac{r^2_{\text{no hemi}}}{r^2} = .78$, and the fraction that owes to half-sites is approximately .22.

As a further test of the two-mode model, diffusion coefficients for each segment were calculated from simulated data of a protein undergoing a random walk on the two-mode landscape. The simulations were implemented using the Gillespie algorithm[64] and data from the simulations were analyzed using identical procedures as those used for determining experimental Ds. The simulations provide an independent verification of our ability to separate the effects of DNA fluctuations and drift from estimates of diffusion coefficients of p53 on segments of λ DNA (*Appendix*, Figure A2). The simulated and experimental diffusion coefficients across the segments correlate strongly ($r = .834$, $p = .004$) (Figure 5.7E). Statistical significance was determined by performing simulations based on the same 500 control landscapes described above. Simulations thus provide similar validation as do analytical results of the sequence-specific sliding of p53 by the mechanism of two modes of interaction with DNA.

5.3 Discussion

We have previously proposed a two-mode model of protein-DNA interaction that allows for fast search and specific binding [39, 37]. Our earlier single-molecule measurements

of p53 sliding on DNA revealed that the protein slides with sufficiently low friction to satisfy the model's requirements for efficient search[43]. The present study shows that p53 can read the sequence of the DNA on which it is sliding, which is essential for sliding to be functional in accelerating target localization. Our data further suggest that the protein reads, in addition to canonical and near-canonical 20-bp full-sites, half-sites of 10 bp. Hemi-specific binding has recently been shown to play a role in transcriptional activation at high p53 expression levels[92], and so it is fitting that p53 should recognize half-sites as well as full-sites while sliding on DNA.

Our results indicate that hemi-specific binding is a general phenomenon of p53-DNA interactions, and not limited to a few known half-site response elements (REs). In addition to transcriptional activation, we conjecture that hemi-specific binding might serve to titrate p53 or bias the pre-activation distribution of p53 on DNA. This latter function especially is suggested by the clustering of degenerate p53 REs has been found near canonical REs, which has also been found for other mammalian TFs[130]. Binding sites for p53 have been identified that contain an odd number of half-sites[96, 95]; hemi-specific binding would allow finer tuning of transcriptional activation of p53's targets.

Our two-mode model of p53-DNA interaction, including hemi-specific binding, is based on a half-site position weight matrix approximation. Although we did take into account variable spacing between half-sites (*Materials and Methods*, Equation 5.6), a number of deviations from our model have been observed. We assume that the contributions of the component half-sites to the full-site binding energy are equally important. It has been found, however, that the first half-site is more conserved among known p53 response elements (REs) than is the second half-site[96] that, according to some characterizations of the p53 RE, the inner 10 bp of a 20-bp full-site more strongly predict binding affinity than the outer 10 bp [131], and that positions 3 and 5 out of 20 are particularly important as

well [98]. Additional deviations from our approximation found in known p53 REs include spacers within half-sites[96] and transcriptional activation from three-quarter sites[92, 132], and differential effects on transcriptional activation of mutating the first versus the second half-site in a full-site consisting of two identical half-sites[133].

Our model of p53-DNA binding energy is based on a PWM approximation that was shown to be sound for the four eukaryotic TFs studied[129], but it omits some observed peculiarities of p53 REs such as gaps within half-sites[96], stronger conservation within a full-site of the first half-site than the second[96], and transcriptional activation from three-quarter sites[132]. Accounting for these complexities might yield a stronger correlation between predicted and experimental D, at the expense of model simplicity.

The two-mode model discussed in here can be generalized into to include transition states or a reaction coordinate of the conformational transition in the protein-DNA complex [39, 37]. Molecular dynamics studies of TF-DNA association indeed show a range of conformations [134]. Since our estimate of D/D_0 is equivalent to the ratio of partition functions of a totally flat landscape and the predicted "golf-course" landscape, the rates of transition between **R** and **S** modes play no role. On sufficiently long timescales of sliding, a protein's diffusivity will be independent of these rates, since keeping the binding energy of the protein in **R** mode and in **S** mode at a given position the same requires that the **R**-to-**S** rate and the **S**-to-**R** rate vary by the same constant factor. A visit to the **R** state that lasts n times longer will happen n times less frequently.

We report here the observation of sequence-dependent 1D diffusional kinetics of a protein on DNA. We offer additional experimental support for the importance of 1D diffusion in the kinetics of transcriptional regulation and protein-DNA recognition. With p53 at least, a full understanding of how its complex promoter architecture functions in transcriptional regulation requires consideration of moves by the protein on DNA even

after it has found its cognate site and the ability of the protein to recognize both full- and half-sites while undergoing those moves. Evidence for a multi-mode model of p53's binding to DNA suggests that the protein's function may be disrupted not only by the comparatively well-studied mutations in residues participating in cognate-site binding, but also by mutations that affect its non-specific interaction with DNA or its ability to transition between specific and non-specific modes, with potential importance for human health.

5.4 Materials and Methods

5.4.1 Materials and data acquisition

The optical setup, DNA constructs, labeled p53, and flow cells (Figure 4.2) were as described in Chapter 4 and an earlier paper [43], with the exception that the protein was labeled with AlexaFluor 555 (Invitrogen) and illuminated with the frequency-doubled 532-nm line of a Nd:YAG laser, an oxygen-scavenging system was used, and fiduciary beads employed to align movies of proteins with movies of DNA (Figure 5.8A).

The concentration of DNA used in this work needed to be significantly lower than in the work of Chapter 4, since the DNA could not be so dense as to prevent us from assigning protein particles to a distinct DNA molecule. This assignment was necessary in order to place trajectories correctly on the contour of the λ DNA. In the previous work, stained DNA illuminated nearly the entire field of view; in the current work, DNA concentration was lowered to approximately 40 DNA molecules per field of view (Figure 5.8B).

The concentrations of protein and salt also had to be altered. To increase the efficiency of data collection, a protein concentration (150 pM) was selected to give an average of one labeled protein per DNA molecule. As the labeling efficiency of for the protein we used was 30%, this resulted in an average of ~3 particles per DNA molecule. As the average

span of DNA covered by a sliding protein ($\bar{n}$ in the language of Part I) was $<$ 3kb and λ DNA is ~50kb long, we did not consider the possibility of proteins "jamming" each other to be a significant concern. A salt concentration of 163 mM was used, rather than the maximum of 125 mM in Chapter 4, both for physiological accuracy and to refresh the DNA with new proteins, which prevented the ratio of bleached to unbleached labels from declining too rapidly.

At the time of the experiments, the amount of data thought necessary to be collected for adequate statistics was such that being able to run longer experiments with longer trajectories would have been a substantial boon. To that end, we explored the use of an oxygen-scavenging system to reduce photobleaching. The system consisted of glucose oxidase (GOx), catalase, and glucose. The GOx catalyzes the reacting of dioxygen with glucose, producing hydrogen peroxide, and the catalase catalyzes the decomposition of two hydrogen peroxide molecules to two molecules of water and one of dioxygen. We found that the oxygen-scavenging system reduced the photobleaching rate by a factor of 2.50, and did not affect the binding and sliding of p53 to DNA. We also experimented with using the antioxidant Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid), but found it did not improve the photobleaching in our system.

Before labeled p53 was introduced to the flow cell, a 0.016% suspension of biotinylated fiduciary beads were flowed in with a concentration and incubation time such that 2 5 beads appeared in each field of view. Bead preparation follows Elenko [135]. When the flow cell had been studded with beads, movies were taken of p53 sliding on flow-stretched λ-phage DNA, with a flow rate of 100μL/min through a flow cell 2mm wide, 100μm tall, and 36 mm long. p53 sliding buffer consisted of 20 mM HEPES (equilibrated to pH 7.9 with NaOH), 150 mM KCl, 0.5 mM EDTA, 2 mM MgCl2, 0.25 mg/mL BSA, and 2.5 mM DTT. p53 concentrations were between 50 and 150 pM. At the end of the experiment, DNA

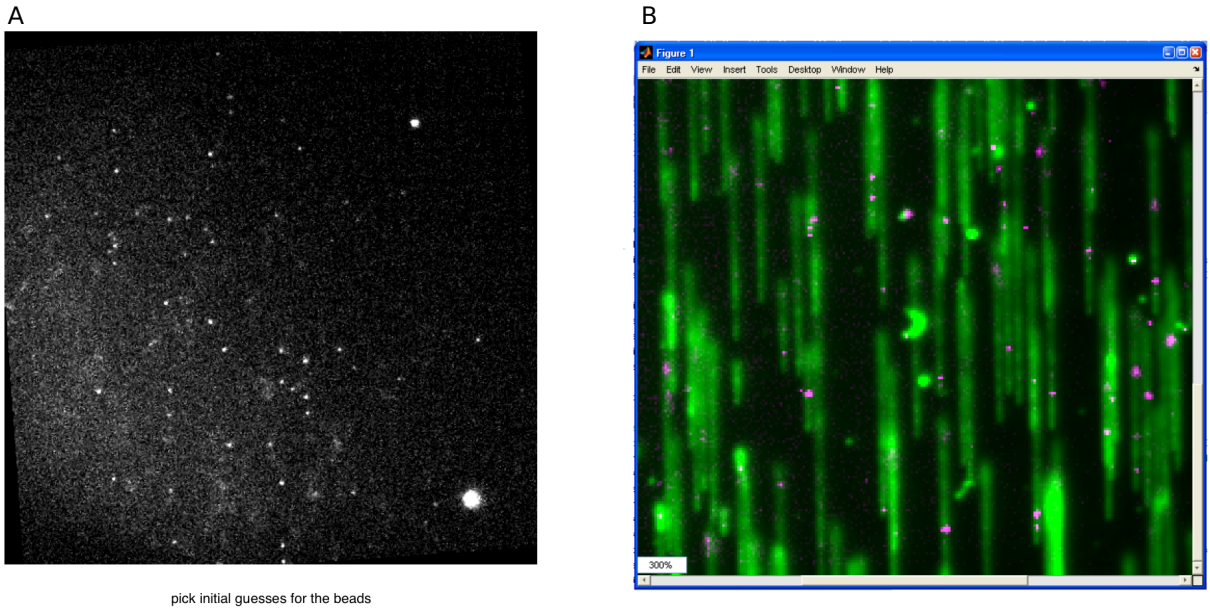


Figure 5.8: Screenshots from a GUI written to facilitate aligning protein and DNA movies. (**A**) Flattened protein movie showing two beads, one in the lower-right, the other in the upper-right. Owing to uneven illumination within the field of view, the beads appear to be of different sizes or brightnesses. After the user clicks on the beads, a similar image appears, this time of the flattened DNA movie. The user clicks on the same beads, and the software aligns the movies and creates a false-color image, (**B**), that superimposes the flattened protein movie (magenta) and the flattened DNA movie (green). The user now selects areas of the movie in which particles are found and to be tracked.

was stained with Sytox Orange (Invitrogen) to show the position of the tethered DNA. The beads allowed movies of proteins sliding and movies of the stained DNA to be aligned despite stage drift (Figure 5.8).

The alignment was aided by the use of software written in MATLAB (Mathworks). We implemented a graphical user interface (GUI) that asked the user to make initial guesses for the bead centers in flattened stacks of the protein and DNA movies (the former shown in Figure 5.8A), and then fit the intensity from the beads to two-dimensional Gaussian functions. The software then translated the protein movies so as to minimize the sum-of-squares error between centers of the Gaussians in the respective movies, and presented

the user with a composite image, whereupon the user could match a particle to track with the tether point of the DNA the particle was located on. Since the DNA was visualized using an intercalating stain at a concentration such that the average distance between stain molecules, approximately 20 bp or 7 nm, would more than an order of magnitude smaller than the width of the point spread function in our optical system, we considered the tether point to be located where the intensity of the stain was midway between the backround intensity and the average intensity along the DNA polymer.

5.4.2 Data analysis

Protein molecules were assigned to individual DNA molecules and their trajectories recorded using scripts written in MATLAB. Positions of the p53 molecules in space along the DNA image were mapped to positions on the contour of the DNA. To achieve this, Brownian dynamics simulations of DNA as a tethered polymer in shear flow were performed to determine the degree of compression in the DNA as a function of the distance along the contour from the tether[136]. Integrating and inverting this function yields a function that transforms positions in the recorded images to positions along the contour of DNA (Figure 5.9). Figure 5.2B shows three sample trajectories.

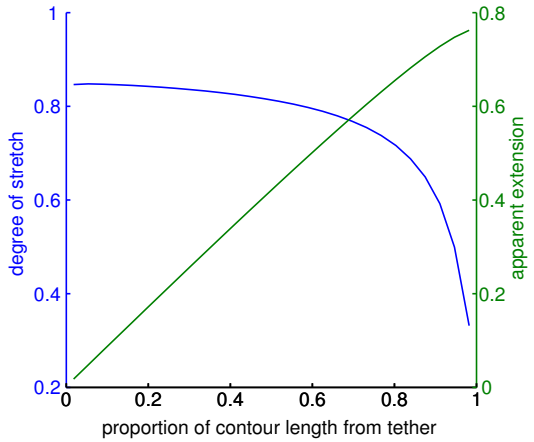


Figure 5.9: Obtaining a microscope-to-contour map of λ-phage DNA. Brownian dynamics simulations were performed to obtain the blue trace, the degree of stretch in the polymer as a function of the distance along the contour from the tether. Integrating this curve gives the green trace, which is a one-to-one map of position on contour to apparent position in our images.

We determined a diffusion coefficient D for each p53 particle using maximum

likelihood estimation, correcting for biased drift owing to buffer flow as well as for fluctuation in the λ-phage DNA on which the proteins diffused. We found D for a particle to be estimated by:

$$D = \frac{1}{2}\Big(\frac{1}{n}\sum_{i}^{n} \frac{(\Delta x_i - v\Delta t_i)^2}{\Delta t_i} - \frac{1}{n}\sum_{\Delta t} n_{\Delta t}\frac{\langle \Delta x_{d,\Delta t}^2 \rangle}{\Delta t} \Big) \tag{5.2}$$

where v is the drift velocity, given by

$$v = \frac{\sum_{j}^{\text{all traj.}} x_{j,final} - x_{j,initial}}{\sum_{j}^{\text{all traj.}} t_{j,final} - t_{j,initial}} \tag{5.3}$$

An N-frame trajectory contains $(N-1)(N-2)/2 \equiv n$ displacements. The ith observed displacements in space and in time are respectively Δx_i and Δt_i. The second sum in Equation 5.2 is over time windows Δt ranging in duration from the camera frame rate, 30ms, to 2s. The quantity under the sum is the mean squared displacement of the DNA itself owing to Brownian fluctuations in the buffer $\langle x_{d,\Delta t}^2 \rangle$ on a timescale of Δt, as calculated from measurements of quantum dots covalently attached to the DNA (*Appendix* 5.A3), divided by Δt, and weighted by the number of displacements with a corresponding Δt. The first sum represents the apparent diffusion coefficient of p53, corrected for drift. Equation 5.2 is derived in Appendix 5.A1.

Once a p53 particle's diffusion coefficient had been determined, the diffusion coefficient was assigned to every midpoint of the particle's trajectory's displacements. Data from the third of the DNA farthest from the tether was discarded owing to the large amplitude of DNA fluctuations beyond that point. The DNA was divided into segments with a width chosen equal to the mean end-to-end distance of remaining trajectories, approximately 2.9 kb. The mean of the diffusion coefficients assigned to positions with in each segment was

calculated and then compared with the predicted diffusion coefficient based on theoretical energy landscapes.

A number of alternative methods of data analysis were pursued but ultimately rejected. With regard to defining which displacements we would use to estimate the D for each segment, at first, we considered only frame-to-frame displacements, and later, we divided individual trajectories into portions of them that stayed within segment boundaries. Once the displacements to use were identified, we used various methods of parameter estimation that were less rigorous than the MLE-based approach described above and in *Appendix* 5.A1. These data analytical methods are described in *Appendices* 5.A4 and 5.A5.

5.4.3 Prediction of diffusion coefficients

To predict diffusion coefficients for each segment, we first built a predicted effective energy landscape $U(x)$, and then calculated the predicted slow-down in each segment based on the landscape. $U(x)$ is based on two component landscapes, one from binding in the **R** mode and the other from binding in a zero-variance ($\sigma_S = 0 k_B T$) **S** mode. In the **R** mode, the protein's binding energy, E_R, is dependent on its position on DNA, x, and in the **S** mode, its binding energy, E_S, is constant. Additionally, in the case of p53, the protein is a dimer of dimers, with each dimer having been shown to be able to bind independently to a 10-bp half-site[91]. For binding in recognition mode, then, the left dimer binds with energy $E_R(x)$, and the right dimer binds with energy $E_R(x + \Delta)$, with Δ the separation in bp between the two dimers.

To determine $E_R(x)$, we scored the λ genome with a position weight matrix (PWM) of p53 half-sites derived from a catalogue of p53 binding sites assembled by Horvath *et al.* [97]. We assume that the differences between scores are proportional to differences between corresponding half-site energies, which are in units of $k_B T$:

$$E_R(x) - E_S = c\,(PWM(x) - PWM_{\text{reference}}) \tag{5.4}$$

$PWM(x)$ is the score for position x, and $PWM_{\text{reference}}$ is the score corresponding to binding energy in the **S** mode. Thus, in the event that a site scores equal to the reference score, the specific and non-specific binding energies for p53 to that site will be equal. We chose a value for $PWM_{\text{reference}}$ based on studies of eukaryotic transcription factor binding energies on defective versions of their consensus sequences [129]. It was observed that for all the transcription factors studied, binding weakened as the consensus sites were mutated to contain one and then two mismatches (equivalent to four bits), but then became no weaker with further mutations. We therefore chose a non-specific reference score equal to the score of the best-scoring half-site minus four bits. Varying $PWM_{\text{reference}}$ by a bit in either direction had little effect on our results. The choice of a four-bit threshold receives some additional justification from FRAP measurements of p53 and two other eukaryotic transcription factors that found all three TFs' search dynamics to be similar[137].

The remaining unknown in Equation 5.4 is the proportionality constant c that relates score to energy. Dissociation constants for p53 binding to the left-hand Mdm2 half-site as well as to random DNA are available from biochemical measurements [88]. At our experimental conditions, p53 favors the Mdm2 half-site by a factor of 47 [88][1], and so for this half-site, we estimate $E_R(x) - E_S = \log(47)k_BT = 3.9k_BT$. Substituting this value into the left-hand side of Equation 5.4, and the site's PWM score minus $PWM_{\text{reference}}$ into the right-hand side gives a value for c of $0.97k_BT$/nat or $0.67k_BT$/bit.

At any site x, the protein may bind in four distinct modes owing to the left and

[1]Experiments whose measurements of p53's affinity for full-, half-, and random sites we used to parametrize our model were performed on 30-bp oligonucleotides whose central 20 bp consisted of the specific sites. If the protein can bind non-specifically off-center, then the true preference for specific sites will be greater than it would appear—see *Appendix* 5.A2.

right dimeres being able each to bind in either mode: (1) both dimers in **S**; (2) left dimer in **S**, right dimer in **R**; (3) left dimer in **R**, right dimer in **S**; and (4) both dimers in **R** (Figure 5.5). The statistical weight of a site x is thus the sum of the Boltzmann factors corresponding to each of the four modes:

$$w(x) = e^{-2E_S} + e^{-(E_S+E_R(x+\Delta))} + e^{-(E_R(x)+E_S)} + e^{-(E_R(x)+E_R(x+\Delta)+\epsilon)} \tag{5.5}$$

The constant ϵ is a cooperativity term representing additional binding energy when both dimers are bound in specific mode. Its value was determined from Equation 5.5 by substituting in energies for the left-hand and right-hand sites of the Mdm2 promoter as determined by Equation 5.4 and our PWM scoring, and substituting experimental values for the K_d of the full Mdm2 site relative to the K_d for a random sequence. From this, we find $\epsilon = -1.39k_BT$, the negative sign indicating that the energy of a protein on a full-site that binds both component half-sites in specific mode is $1.39k_BT$ lower than it would be absent any cooperativity.

A small (~10%) proportion of known p53-binding sites include a gap of 1-14 bp between half-sites. To allow gapped full-sites to be treated as such in our predicted energy landscape, $E_R(x+\Delta)$ at each binding site was assigned as:

$$E_R(x+\Delta) = \min_i \left(E_R(x+\Delta_0+i) - c\,\log(f_i/f_0)\right); \quad i = 0, ..., 14 \tag{5.6}$$

Δ_0 is the length of a half-site, 10bp, and thus the separation between half-site start positions in the absence of a gap. The index i is over gaps of length 0 to 14, and f_i is the frequency of gaps of length i in the dataset used to build the PWM. The second term under the minimum accounts for the suboptimal binding conformation the protein must adopt when

binding to half-sites separated by a gap. As $f_{i>0} < f_0$, gapped full-sites suffer an energy penalty, while full-sites with zero gap suffer none.

Setting the energy scale such that $E_S \equiv 0$, Equation 5.5 becomes

$$w(x) = 1 + e^{-E_R(x+\Delta)} + e^{-E_R(x)} + e^{-(E_R(x)+E_R(x+\Delta)+\epsilon)} \tag{5.7}$$

A single-mode model would not include non-specific binding and thus omit all but the final term in Equation 5.7, and a model that disallowed hemi-specific binding would omit the middle two terms. From this function of the statistical weights across all positions, we may treat p53 as interacting with DNA on a "golf-course landscape", the energy at position x of which is equal to the negative logarithm of $w(x)$.

$$U(x) = -\log w(x) \tag{5.8}$$

We used the resulting effective landscape to calculate D_{theo}. We segmented the landscape at the same positions as we did the experimental data, and for each segment predicted the diminution in diffusion coefficient owing to sequence-specific binding by estimating the mean ratio of the time during a visit to the segment that the protein spends sliding on DNA, t_s, versus the total time that it spends on DNA.

$$\frac{D}{D_0} = \left\langle \frac{\Delta x^2/2t_{total}}{\Delta x^2/2t_s} \right\rangle = \left\langle \frac{t_s}{t_{total}} \right\rangle \tag{5.9}$$

D_0 is diffusion coefficient in the absence of sequence-specific binding, i.e., D on a completely smooth landscape, without an $\mathbf{R}$ mode. The ratio t_s/t_{total} for a trajectory $\mathbf{x}$ is

$$\frac{t_s}{t_{total}} = \frac{\sum_i^{\mathbf{x}} \exp(-2E_S)}{\sum_i^{\mathbf{x}} \exp(-U(x_i))} \tag{5.10}$$

where $U(x_i)$ is the effective energy at site x_i, which is the ith site visited in trajectory $\mathbf{x}$. If the transition state for translocating between two sites is constant across all sites–equivalent to assuming that for any position x on DNA, p53's microscopic step rates to positions $x-1$ and $x+1$ are equal or that traps are isolated–then averaging over trajectories results in a uniform distribution of visits to all sites in a given segment, and

$$\left\langle \frac{t_s}{t_{total}} \right\rangle = \frac{n \exp(-2E_S)}{\sum_x^n \exp(-U(x))} \tag{5.11}$$

where n the number of sites in the segment. The right-hand side of Equation 5.11 consists entirely of constants, and E_S is defined to be zero, so

$$\frac{D}{D_0} = \frac{1}{\frac{1}{n}\sum_x^n \exp(-U(x))} \tag{5.12}$$

that is, the diffusion coefficient is diminished by a factor equal to the average of e raised to the effective energy in the segment. Since p53's half-site-binding sequence logo is not perfectly palindromic, $\exp(-U(x))$ was taken to be the mean for the forward and reverse strands.

Once we computed D/D_0 for each segment, we made a correction to account for uncertainty in the assignment of experimental displacements to segments owing to DNA fluctuations, described in *Appendix* 5.A3. We then assessed the quality of our predicted diffusion coefficients by computing Pearson's correlation coefficient r_{expt} between experimental diffusion coefficients D_{expt} and predicted D/D_0 over the segments. To assess the statistical significance of r_{expt}, we constructed 500 control landscapes by randomly permuting the λ-genome scores, giving rise to a permuted $E_R(x)$. Owing to the 10-bp half-site PWM having the bulk of its information content in two nucleotides three positions apart, permuting the PWM is not a viable control, as $10-3=7$ out of $10^2=100$ permuted PWMs will closely resemble the original PWM. To obtain p-values, we calculated a correlation coefficient r_{ctl}

between each control landscape's predicted $\frac{D}{D_0}$ and D_{expt} over the segments of λ DNA, and determined the fraction of r_{ctl} greater than or equal to r_{expt}.

Before we had concluded that the diminution in diffusion coefficient should be given by Equation 5.12, we compared experimental diffusion coefficients for a segment with the reciprocal of the variance in energies in that segment, and computed the Spearman's correlation coefficient ρ_{expt}. The correlation was strong ($\rho = .817$, $p < .005$), but we were unsatisfied with lacking a fully theorized prediction of D_{expt} and having to resort to ranked correlation.

5.4.4 Simulations

We simulated random walks on the predicted and control landscapes using the Gillespie algorithm[64]. At first, when we were using less sophisticated and powerful data analysis (*Appendices* 5.A4 and 5.A5), we sought to estimate how much data we would need to collect to identify p53 slowing down on individual predicted energy minima. As a test, we simulated a protein undergoing a random walk on a landscape that was totally flat except for traps or clusters of traps of 4 or 5 k_BT, which we estimated to be the depth of the predicted full-sites we found on lambda using a PWM. We found that even six months of data collection would not be enough to reliably identify individual traps (Figure 5.10). This motivated attempts at making a DNA construct suitable for single-molecule microscopy that would have deeper traps, consisting of the strongest known p53 binding site (Chapter 6).

In simulations on landscapes that were nearly entirely flat, we considered the possibility of multiple proteins on the same DNA molecule, which potentially could jam each other. For the sake of speed, proteins made 100 steps at once, unless they were within 100 steps of another protein, in which case a "mini-simulation" would be entered, in which steps were made one at a time until the proteins were no longer close to each other.

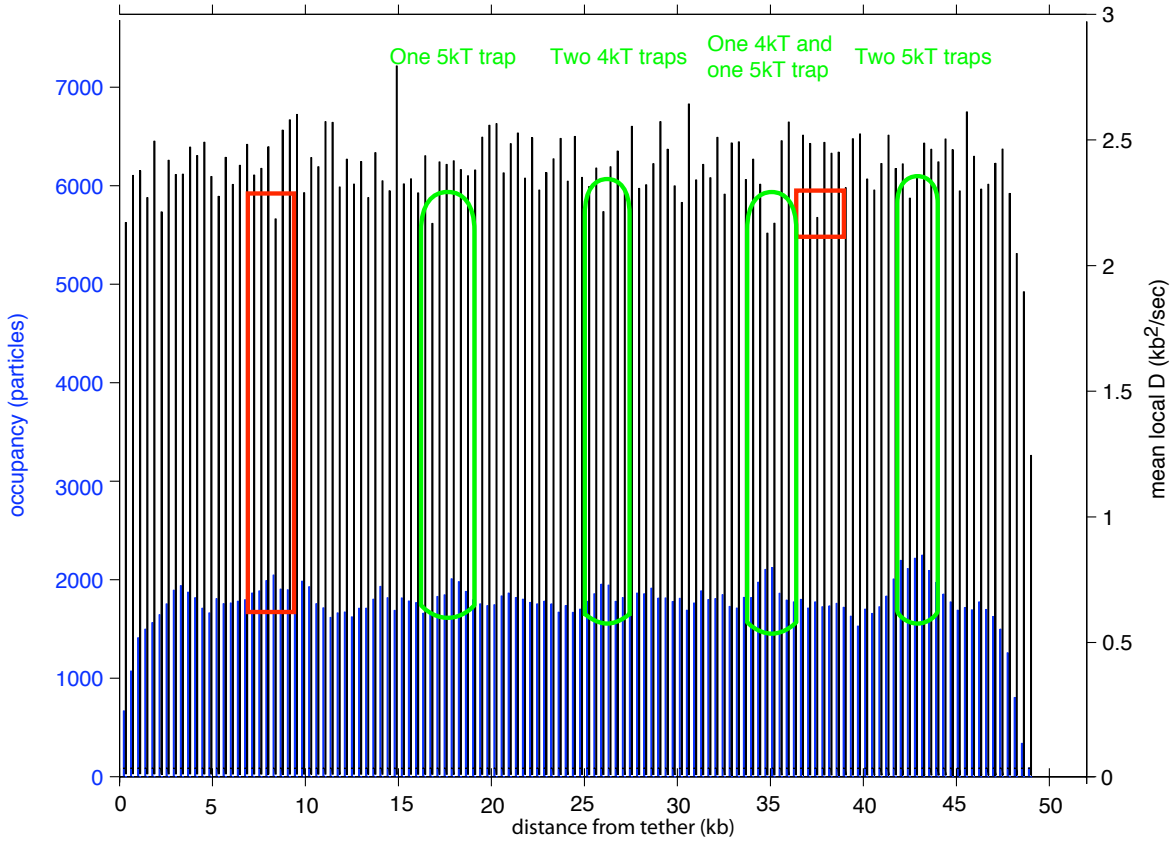


Figure 5.10: Simulations on a flat landscape with a few traps. Blue bars are the number of occupancy events in each segment; black bars are the mean diffusion coefficients in the segment based on frame-to-frame displacements. Green "bacilli" indicate where traps were placed and the corresponding peaks in occupancy and troughs in D. Red rectangles indicate spurious traps—segments that appear from the data to contain a trap but in fact do not. Simulated data was estimated to represent six months of actual data collection.

The mini-simulation was also used when the proteins were near traps. When simulations on the predicted effective landscape $U(x)$ were performed, the simulations were rewritten without the mini-simulation and with one step at a time, as the probabilities of stepping left or right on the effective landscape vary arbitrarily; *i.e.* nearly every position is a "trap". Additionally in the interest of speed, statistical weights for each move were precomputed.

Noise corresponding to DNA fluctuations was added to the simulated trajectories (*Appendix* 5.A3), and then the simulated data were treated identically to the experimental data, and their correlation with experiment and theory determined.

5.5 Acknowledgements

The labeled protein was provided by F.H. The apparatus was set up by Drs. Joseph Loparo and Candice Etson, who also helped in troubleshooting and maintaining it. J.L. collected the data and optimized the experimental conditions, with helpful discussions and technical advice from A.T., Anna Kochaniak, Drs. Loparo and Etson, and Dr. Mark Elenko. Quantum-dot movies were taken by A.T. Nearly all of the data-analysis software was written and designed by J.L.; initial scripts were adapted from ones written by A.T. and Dr. Etson, with helpful discussions with Drs. Elenko and Etson. W.U. performed the Brownian dynamics simulations. A.T., Dr. Etson, Geoff Fudenberg, Maxim Imakaev, and Christopher McFarland participated in helpful discussions in the course of the data analysis. J.L. created the figures and wrote the manuscript from which much of the material in this chapter is taken.

5.A Appendix

5.A1 Derivation of MLE diffusion coefficients

For every p53 particle, a drift rate, v, and diffusion coefficient, D, were determined using maximum likelihood estimation (MLE). Assuming that a particle's displacement due to drift is independent of its displacement due to diffusion, and that the particle's displacements are all independent, the MLEs for a particle's v and D in the absence of DNA fluctuations are derived as follows.

$$\begin{aligned} p(\Delta x; v, D) &= \exp\left(\frac{-(\Delta x - v\Delta t)^2}{4D\Delta t}\right)\left(4\pi D\Delta t\right)^{-1/2} \\ L(\Delta x_1, ..., x_n | v, D) &= \exp\left(\sum_i^n \frac{-(\Delta x_{i,p} - v\Delta t_i)^2}{4D\Delta t_i}\right)\prod_i^n \left(4\pi D\Delta t_i\right)^{-1/2} \\ \log L &= -\sum_i^n \frac{(\Delta x_{i,p} - v\Delta t_i)^2}{4D\Delta t_i} - \frac{1}{2}\sum_i^n \log(4\pi D\Delta t_i) \end{aligned} \tag{5.A1}$$

$\Delta x_{i,p}$ is displacement i of the protein on DNA, which takes place over the duration Δt_i. Taking the partial derivative of L with respect to the drift rate, v, and setting the result equal to zero,

$$\begin{aligned} 0 = \frac{\partial \log L}{\partial v} &= \sum_i^n \frac{2\Delta x_{i,p}\Delta t_i - 2v\Delta t_i^2}{4D\Delta t_i} \\ &= \sum_i^n \left(\Delta x_{i,p} - v\Delta t_i\right) \\ v &= \frac{\sum_i^n \Delta x_{i,p}}{\sum_i^n \Delta t_i} \end{aligned} \tag{5.A2}$$

Here, the index i is over the largest non-overlapping set of $\frac{\Delta x_{i,p}}{\Delta t_i}$, which are the final and initial frames of each trajectory j, so:

$$v = \frac{\sum_{j}^{\text{all traj.}} x_{j,final} - x_{j,initial}}{\sum_{j}^{\text{all traj.}} t_{j,final} - t_{j,initial}} \tag{5.A3}$$

We now take the partial derivative with respect to the diffusion coefficient, D, and equate to zero:

$$\begin{aligned} 0 = \frac{\partial \log L}{\partial D} &= \sum_{i}^{n} \frac{(\Delta x_{i,p} - v\Delta t_i)^2}{4D^2 \Delta t_i} - \frac{1}{2}\sum_{i}^{n} \frac{1}{D} \\ &= \sum_{i}^{n} \frac{(\Delta x_{i,p} - v\Delta t_i)^2}{\Delta t_i} - 2nD \\ D &= \frac{1}{2}\frac{1}{n}\sum_{i}^{n} \frac{(\Delta x_{i,p} - v\Delta t_i)^2}{\Delta t_i} \\ D &= \frac{1}{2}\frac{1}{n}\sum_{i}^{n} \frac{\Delta x_{i,p}^2 - 2\Delta x_{i,p} v\Delta t_i + v^2 \Delta t_i^2}{\Delta t_i} \end{aligned} \tag{5.A4}$$

The observed displacements, Δx_i, are in fact the sum of displacement from protein diffusion, $\Delta x_{i,p}$, and displacement from DNA fluctuations, $\Delta x_{i,d}$. Substituting $\Delta x_{i,p}$ with $\Delta x_i - \Delta x_{i,d}$ in Equations 5.A2 and 5.A4, and substituting $\Delta x_{i,p}^2$ with $\Delta x_i^2 - 2\Delta x_{i,p} - \Delta x_{i,b}^2$ in 5.A4, yields the following:

$$v = \frac{\sum_i^n \Delta x_i - \Delta x_{i,d}}{\sum_i^n \Delta t_i} \tag{5.A5}$$

and

$$D = \frac{1}{2}\frac{1}{n}\sum_{i}^{n} \frac{\Delta x_i^2 - 2\Delta x_{p,i}\Delta x_{d,i} - \Delta x_{d,i}^2 - 2\Delta x_i b\Delta t_i + 2\Delta x_{d,i} v\Delta t_i + v^2\Delta t_i^2}{\Delta t_i} \tag{5.A6}$$

Separating the terms under the sum in the expression for v gives

$$v = \frac{\sum_i^n \Delta x_i}{\sum_i^n \Delta t_i} - \frac{\sum_i^n \Delta x_{i,d}}{\sum_i^n \Delta t_i} \tag{5.A7}$$

The second term in Equation 5.A7 vanishes because the displacements due to DNA fluctuations, $\Delta x_{d,i}$, have mean zero, and so the drift rate is simply that given in Equation 5.A2. In the equation for D (5.A6), the DNA displacements are likewise independent of the protein displacements, $\Delta x_{p,i}$, and the drift, $b\Delta t_i$, so the sums of the cross terms $2\Delta x_{p,i}\Delta x_{d,i}$ and $2\Delta x_{d,i} v\Delta t$ also go to zero. Eliminating these terms and separating into four remaining sums yields

$$D = \frac{1}{2}\Big(\frac{1}{n}\sum_i^n \frac{\Delta x_i^2}{\Delta t_i} - \frac{1}{n}\sum_i^n \frac{\Delta x_{i,d}^2}{\Delta t_i} - \frac{1}{n}\sum_i^n \frac{2\Delta x_i v\Delta t_i}{\Delta t_i} + \frac{1}{n}\sum_i^n \frac{v^2\Delta t_i^2}{\Delta t_i}\Big) \tag{5.A8}$$

This is equivalent to Equation 5.2. The third and fourth terms are known from the estimate of v in Equation 5.A2 and from observed Δx_i and Δt_i. The second term in Equation 5.A8 is equivalent to

$$\frac{1}{2}\frac{1}{n}\sum_i^n \frac{\Delta x_{i,d}^2}{\Delta t_i} = \frac{1}{2}\frac{1}{n}\sum_{\Delta t} n_{\Delta t}\frac{\langle \Delta x_d^2\rangle}{\Delta t} \tag{5.A9}$$

where $n_{\Delta t}$ is the number of displacements with duration Δt in the trajectory, and Δx_d are displacements owing to DNA fluctuation. Trajectories of DNA fluctuations were measured in previous work[53] by examining the trajectories of quantum dots covalently attached to λ-phage DNA at known positions. The expression in Equation 5.A9 is thus the expected contribution of DNA fluctuations to the apparent diffusion of the protein (*Appendix* 5.A3).

5.A2 Non-specific binding in model parametrization

To parametrize our scored λ genome into an energy landscape, we used dissociation constants from *in vitro* affinity assays of p53 and 30-bp oligonucleotides bearing full-sites, half-sites, and random DNA [88]. Since p53's binding site is 20-bp long, it is possible that one or more non-cognate sites are available for p53 to bind to on either side of the full-

and half-sites. Indeed, oligonucleotides of only 26 bp have been used to study binding between p53 and its cognate sites[138], so it is not improbable that a 30-bp oligonucleotide can accommodate p53 binding at least four non-cognate sites. If this is the case, then the apparent preference of p53 for half-site 30-mers relative to random 30-mers, of approximately a factor of 8 reflects a true preference for a single half-site over a single random site of 35:

$$\begin{aligned} \frac{n\exp(-E_n)+\exp(-E_h)}{n\exp(-E_n)} &= x_{hn} \\ \frac{\exp(-E_h)}{\exp(-E_n)} &= n(x_{\mathrm{hn}}-1) \end{aligned} \tag{5.A10}$$

where n is the number of sites available on the oligonucleotide for binding, including the cognate site, E_h and E_n are half-site and non-cognate binding energies, respectively, and x_{hn} is the apparent factor by which p53 prefers to bind the half-site in hemi-specific mode relative to non-cognate DNA in non-specific mode. For values of $n = 5$ and $x_h = 8$, the true preference for half-sites is approximately four-and-a-half times greater than the apparent preference, corresponding to an energy difference of $1.5k_BT$.

This energy difference is reflected in a greater value for the proportionality constant c relating the score of a site to its energy. With available binding sites flanking the cognate site, $c = 0.97k_BT/nat$, while with four sites on either side ($n = 5$ in Equation 5.A11), it increases to $1.37k_BT/nat$. This has the concommitant effect of raising the energy of cooperativity between specific-mode binding in the two dimers (that is, raising the energy of the fully-specifically-bound state) from $\epsilon = -1.39k_BT$ to $+0.19k_BT$, that is, specific binding becomes weakly anticooperative. The increase in c amounts to a more rugged landscape, with deeper wells at half- and full-sites, while the decrease in ϵ causes full-site binding to become weaker. The information content of the p53 sequence logo is such that

these two effects are similar in magnitude and opposite in sign, and thus largely cancel each other out. For a pair of adjacent half-sites that each score a typical 4 bits better than the score corresponding to non-specific binding, s_0, the energy for fully-specific binding, which is the dominant form of binding on such a site, equals $2 \cdot (\log(2)\text{nat/bit}) \cdot 4 \text{ bits} \cdot 0.97 k_B T/\text{nat} + 1.39 k_B T = 6.8 k_B T$ in the absence of available flanking sites, and $2 \cdot (\log(2)\text{nat/bit}) \cdot 4 \text{ bits} \cdot 1.37 k_B T/\text{nat} - 0.19 k_B T = 7.4 k_B T$. We presented results assuming no flanking sites, but the landscapes based on the availability of 4 flanking sites are very similar in the predicted local diffusion coefficients they produce: both have a correlation coefficient of .81 with experimental D.

A similar treatment for the true preference of a dimeric DNA-binding protein for binding a full-site in full-specific mode relative to a non-cognate site in non-specific mode, $\exp(-2E_h - \epsilon)/\exp(-E_n)$, as a function of the apparent preferece, denoted x_{fn}, follows:

$$\frac{n\exp(-E_n) + 2\exp(-E_h) + \exp(-2E_h - \epsilon)}{n\exp(-E_n)} = x_{\text{fn}}$$

Rearranging and substituting in Equation 5.A11,

$$\begin{aligned} \frac{n\exp(-E_n) + 2n(x_{hn} - 1)\exp(-E_n) + \exp(-2E_h - \epsilon)}{n\exp(-E_n)} &= x_{\text{fn}} \\ \frac{\exp(-2E_h - \epsilon)}{\exp(-E_n)} &= n(x_{\text{fn}} - 2x_{\text{hn}} + 1) \end{aligned} \quad (5.A11)$$

Although non-specific binding to the oligonucleotides did not turn out to affect our results substantially, this owes to an accident of the parameters relevant to our system. Non-specific binding of proteins to specific probes receives little attention, and yet is necessary to consider when making accurate estimates of binding preferences.

5.A3 Interpolations of DNA-fluctuation variance and distributions

We used our data from earlier work [53] of quantum dots covalently attached to positions on λ-phage DNA one-third and two-thirds the distance from the tether to estimate the estimate the mean apparent diffusivity owing to DNA fluctuations, $\langle \Delta x_d^2 \rangle$, in Equation 5.A9. $\langle \Delta x_d^2 \rangle$ at position x along the contour is expected to fluctuate according to a polynomial in x with non-zero linear and quartic coefficients [79]. For all time windows Δt up to a maximum of two seconds, we fit these coefficients to the observed variance in displacement of the quantum dots at $x = 1/3L$ and $x = 2/3L$ (L = the contour length of λ DNA), and an assumed zero-variance point at the tether, between frames separated by Δt to arrive at an expression for $\langle \Delta x_d (\Delta t)^2 \rangle$:

$$\left\langle \Delta x_d^2(\Delta t) \right\rangle = a_1(\Delta t) \cdot x + a_4(\Delta t) \cdot x^4 \tag{5.A12}$$

The same quantum-dot (QD) data was used to correct estimates of D/D_0 for the uncertainty in the assignment of experimental displacements to segments owing to DNA fluctuations. We determined for each segment's D/D_0 the proportion p of the apparent population of the segment s that can be expected to originate in fact from neighboring segments $s-1$ to the left and $s+1$ to the right (Figure A1, green bars):

$$\frac{D_{\text{corrected}}}{D_0}[s] = (1 - p_{-1} - p_{+1}) \frac{D}{D_0}[s] + p_{-1} \frac{D}{D_0}[s-1] + p_{+1} \frac{D}{D_0}[s+1] \tag{5.A13}$$

$$p_{\Delta s} = \int_{-w/2}^{+w/2} \mathrm{Q}(x|s + \Delta s) * \frac{1}{w} \, \mathrm{d}x \tag{5.A14}$$

The variable s identifies the segment whose D/D_0 is estimated; $p_{\pm 1}$ is the contribution to a segment's observed population of neighboring segments $s = \pm 1$. The integral is over all base pairs in the indicated segment. $\mathrm{Q}(x|s)$ is the distribution of longitudinal DNA

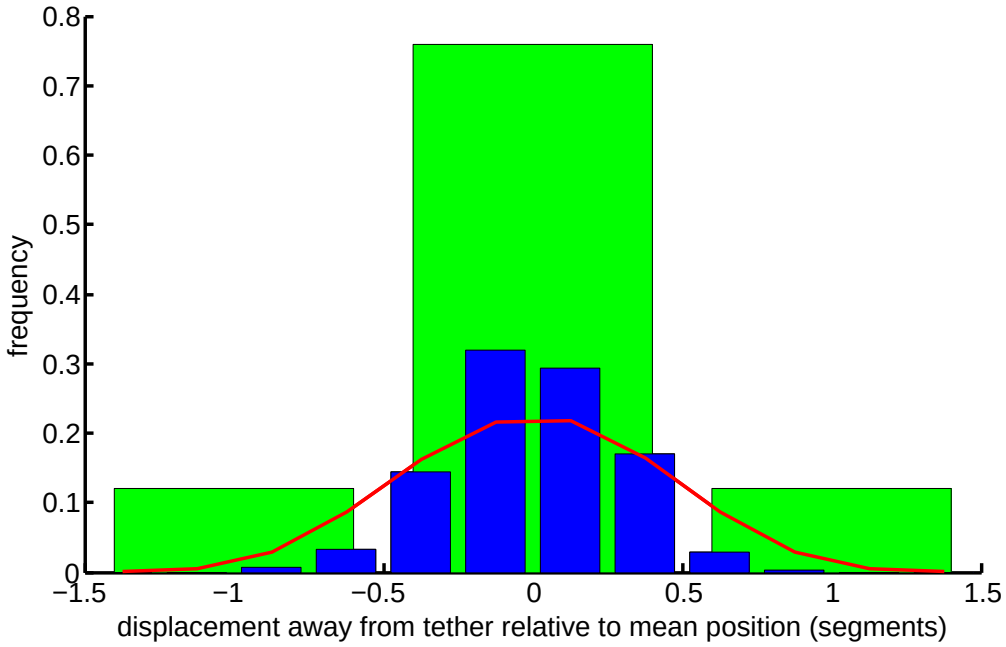


Figure A1: Histograms used to determine p_{Δ_s} and $Q(x|s)$. Data from the quantum dot 1/3 the λ-length from the tether is shown, with $\Delta t = 1$ frame (= 30ms). Blue bars are the distribution of frame-to-frame displacements in quantum-dot positions, Δx_d. Red trace is the blue bars convolved with a uniform distribution one segment wide. Green bars are the red trace binned into segments.

displacements from equilibrium for segment s (Figure A1, blue bars), normalized such that $\int_0^\infty Q(x|s)\mathrm{d}x = 1$, which we obtained from the same quantum dot measurements used to correct experimental D for DNA fluctuations. We assumed that the density of data giving rise to observed diffusion coefficients in each segment was uniform within that segment, and so convolved the distributions of the quantum dots displacements with a uniform distribution the width of a segment, $1/w$ (Figure A1, red trace). It is worth remarking that the distribution of DNA displacements, Q, is itself a function of distance from the tether, so the convolution kernel widens as it moves farther from the tether.

To determine the distribution $Q(x|s)$ used in Equation 5.A14, we constructed sample distributions of the position of the QDs at 1/3 and 2/3 the length of the DNA from the tether, about their mean positions. The variances of these distributions were used to

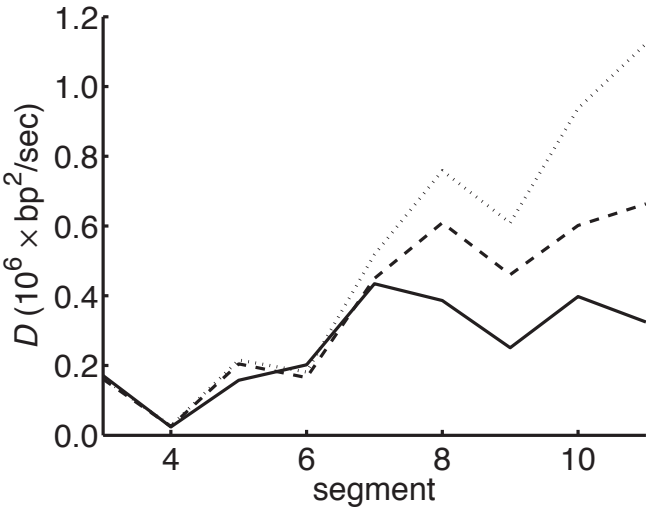


Figure A2: Diffusion coefficient (bp^2/sec)as a function of segment for simulated data. Solid trace is for simulations with no noise added; dotted line is for simulations with added noise of a magnitude and distribution equal to observed displacements of quantum dots attached to λ-phage DNA. Dashed line is from the same noise-added data as dotted line, with apparent diffisivity owing DNA fluctuations subtracted out per Equations 5.2 and 5.A9.

find the coefficients of a similar polynomial as the one in Equation 5.A12. Interpolated distributions consisted of a linear combination of the two closest experimental QD distributions, including a zero-variance delta distribution assumed for the tether point, such that the variance of the interpolated distribution at a position s equaled the fitted polynomial evaluated at that position:

$$Q(x|s) = \begin{cases} b_s\, Q\big(x|0\big) \;+\; (1-b_s)\, Q\big(x|\tfrac{1}{3}L\big) & 0 < s \le \tfrac{1}{3}L \\ b_s\, Q\big(x|\tfrac{1}{3}L\big) \;+\; (1-b_s)\, Q\big(x|\tfrac{2}{3}L\big) & \tfrac{1}{3}L \le s < \tfrac{2}{3}L \end{cases} \tag{5.A15}$$

$$\mathrm{Var}\big(Q(x|s)\big) = a_1 s + a_4 s^4 \tag{5.A16}$$

The QD measurements were also used to add noise to simulations. Results from simulations with no noise and from simulations with this noise added and then subtracted out as described in *Materialis and Methods* 5.4.2 are compared in Figure A2.

5.A4 Alternative data analysis: criteria for selecting displacements

Our first technique for estimating D involved simply taking every frame-to-frame displacement, squaring it, and dividing it by its duration, which in the event of missing frames was larger than the frame rate of 30ms. These square-displacments over time, or diffusivities, were assigned to the segment in which their midpoint was found, and then all the diffusivities in a segment were averaged and halved to arrive at D_{expt}. At the time, the segments used were much smaller, 500bp rather than 2.9kb. It was thought that even though the component of the diffusivity owing to DNA fluctuations, $\frac{\Delta x_d^2}{\Delta t}$, was typically larger than that owing to protein sliding, $\frac{\Delta x_d^2}{\Delta t}$ varied little between adjacent segments, and so local minima should still be visible. We then simulated collecting six months worth of data given the experimental magnitude of DNA fluctuations and determined that even with that much data, true energy minima would not be definitively discernable (Figure 5.10).

The method used in the results presented here for assigning diffusion coefficients, D, to segments is to determine the maximum likelihood estimate of D for every particle j (Equation 5.A4), and then assign D_j to the segment or segments in which the particle is found. A segment's overall diffusion coefficient is average of its D_j's, weighted by the number of displacements in particle j's trajectory within the segment. In earlier data analysis, we did not estimate a D for each particle; rather, to determine a segment's D, we took every displacement Δx and corresponding time window Δt over all the particles with displacements in that segment, and performed parameter estimation in a variety of ways from these very large lists of Δx's and corresponding Δt's (*Appendix* 5.A5).

We dealt with trajectories that crossed segment boundaries in a number of ways. The most straightforward was to divide each trajectory into fragments demarcated by whenever it crossed between segments. Thus, a trajectory j that began in segment s, walked into segment $s+1$, and then returned to segment s would be divided into three entirely

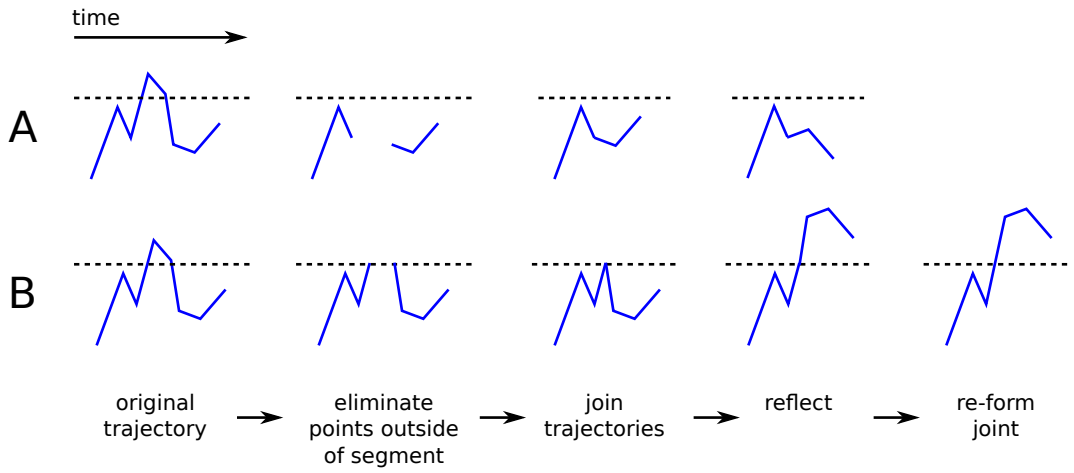


Figure A3: Schematic of alternative data analytical technique. The first row, (**A**) represents the technique of equating, or registering, points of a trajectory separated only by a part of the trajectory that leaves and then returns to a segment. The second row, (**B**) represents the technique of equating points interpolated at segment boundaries. In (**A**), points outside of the segment are removed along with any displacements they constitute. The trajectory fragments as then joined, with the latter trajectory fragment translated in space so that its initial point is aligned with the final point of the former trajectory fragment. The trajectory subsequent to the joining point is reflected in space about the point. In (**B**), a particle's path between points that lie on opposite sides of a segment boundary is interpolated, and the points where the interpolated paths intersect the segment boundary are joined. As with (**A**), the trajectory subsequent to the joining point is then reflected. Finally, the interpolated, joining point is removed.

separate trajectories j_1, j_2, and j_3. A displacement between a point in j_1 and j_3 would not be included in our estimation of segment s's diffusion coefficient. A problem with this method is that, for a given Δt, it is biased against large displacements, as those are more likely to cross segment boundaries.

It was thought that finding a way to stitch together parts of a trajectory that lay in the same segment but were separated by excursions into adjacent segments would allow us to use our data more efficiently. One such way is illustrated in Figure A3A. For all segments that a trajectory visited, points in the trajectory lying outside of the segment were removed, and the initial point of trajectory fragment j_{n+1} would be equated to the final point of the prior trajectory fragment j_n. This required translating fragment j_{n+1} in space. Since the

segment boundaries were treated as reflecting boundaries for a random walk, each time a fragment was joined to the previous one, it and all subsequent fragments were reflected in space about the point at which it was joined to the previous fragment. The first fragment was considered to join a previous dummy fragment so that odd trajectory fragments would be reflected once on net, while even trajectories would be end up unreflected relative to their original orientation.

Both this joining method and the method that did not join trajectory fragments suffer from discounting the regions of segments near the boundaries, as displacements that cover those regions are likely to cross the segment boundary and thus be eliminated. We attempted to counteract this bias with an alternative joining procedure (Figure A3B). Here, the point in time at which a frame-to-frame displacement crosses a segment boundary is interpolated from the surrounding two frames. It is now these points that are joined, and odd-numbered trajectory fragments are reflected about these points. The points are not treated as components of the trajectories, however, and are used only for joining and reflecting purposes. Since the interpolated points are found at arbitrary time points, the Δts for a segment are no longer integer multiples of the frame rate. This property limited the parameter estimation techinques we could apply to it, as will be seen in *Appendix* 5.A5.

5.A5 Alternative data analysis: parameter estimation

We considered a number of methods to estimate local D. A major concern was how to subtract the effect of DNA fluctuations; in all cases drift was straightforward to correct for: we subtracted $v\Delta t_i$ from each displacement i, with v estimated by Equation 5.A3. After correcting for drift, methods to estimate D included:

1. Fitting to a scatterplot of all diffusivities in a segment, $\Delta x^2/\Delta t$.

    - Fit the scatter to a constant (contribution from DNA fluctuations) plus a line

(normal diffusion).

- Fit the scatter to a phenomenological function of DNA fluctuations plus a line.
- Correct each Δx^2 for the expected contribution from DNA fluctuations in advance; fit to a line.

In all cases, the line should have a slope equal to $2D$.

2. For all the Δx's corresponding to a given Δt, fit to a Gaussian distribution. Then take the fit variances as a function of Δt and fit these to a line. The slope of the line estimates $2D$. Since autocorrelation in the DNA fluctuations vanishes for $\Delta t > 90ms = 3$ frames, begin the fit at $\Delta t = 3$ frames. The fluctuations should then add the same variance to each of the Gaussians independent of Δt, and thus not affect the slope of the variances over Δt. Various weighting methods can be employed:

   - The linear fit can be weighted by the number of observations comprising each point. For example, if there are 10,000 displacements with $\Delta t = 10$ frames and 9,500 displacements with $\Delta t = 11$ frames, the point representing the variance in Δx for $\Delta t = 11$ frames receives in linear fit a weight of .95 relative to the point representing the variance for $\Delta t = 10$ frames.
   - The points of the linear fit can be weighted by the quality of the fit of the Gaussian distributions they represent.
   - The points of the linear fit can be weighted by Δt, on the reasoning that longer-time displacements have more inherent averaging.

   Since these methods depend on having distributions of Δx's for a given Δt, they are incompatible with joining method that interpolates displacements at segment boundaries (*Appendix* 5.A4).

3. Normalize displacements by the square-root of the corresponding duration (Equation 5.A18). These normalized displacements Δx^{norm} were then fit to a Gaussian distribution, and the variance of the fit distribution taken as the estimate of $2D$.

$$\Delta x^{\text{norm}} = \left\{ \frac{(x_{i,n} - x_{i,m})}{\sqrt{t_{i,n} - t_{i,m}}} : \; n > m \,; \; i \text{ over all trajectories} \right\} \tag{5.A17}$$

$$\left\{ \Delta x_j^{\text{norm}} \right\} \sim N(\mu, 2D) \,; \;\; j \text{ over all normalized displacements} \tag{5.A18}$$

A variation on this method is to subtract estimated contributions owing to DNA fluctuations from the displacements going into Gaussian fits rather than from estimated fit parameters.

Method 2, while the most *a priori* appealing, turned out to be sensitive to weighting functions and did not enjoy as strong averaging as Method 3. The variation on method 3 resulted in non-Gaussian distributions of displacements and were thus not suitable for a Gaussian fit. All of these methods are at least somewhat *ad hoc*, and indeed Method 3 itself was superseded by the MLE-based method discussed in *Materials and Methods*.

5.A6 Supplemental Figure

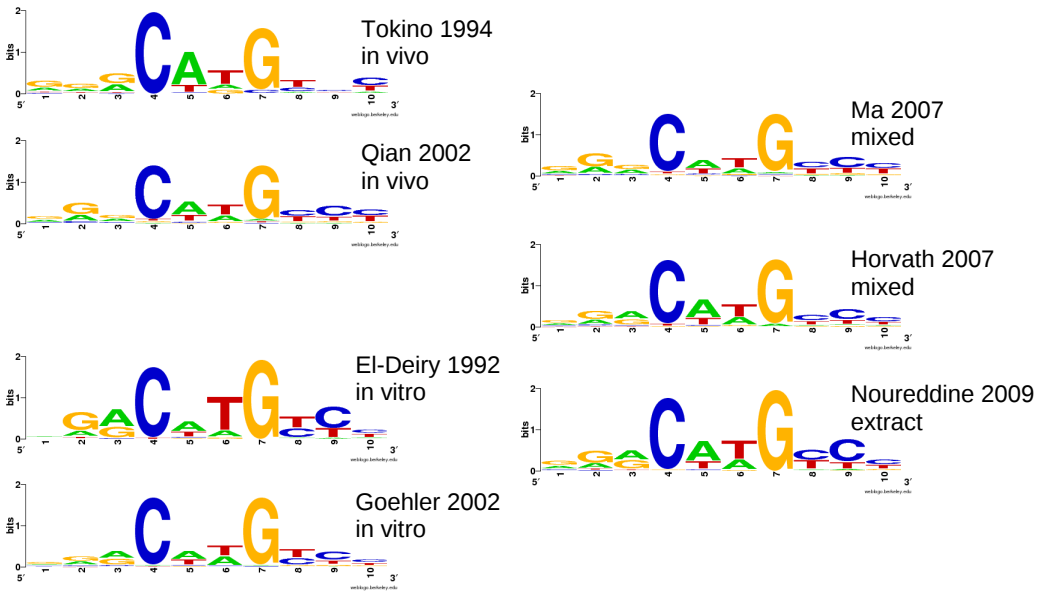


Figure A4: Sequence logos of the p53 half-site from a variety of postition weight matrices [139, 93, 140, 94, 96, 97, 98].

Chapter 6

Implications and future directions

6.1 Experimental improvements

Here, I suggest a number of experimental extensions of the work described in this Part, including one that was pursued but set aside in favor of the work in Chapter 5.

6.1.1 Single-molecule studies of p53 on long DNA with a known target

The studies in Chapters 4 and 5 worked with p53 sliding on λ-phage DNA, a genome p53 has never seen in its evolutionary history. As expected by chance, λ DNA contains a few(~20) sites that our PWM scores to be at least as good as the weakest experimentally verified p53 binding sites. We were interested, however, in visualizing p53's search for a known strong binding site. We hoped to assess whether the protein ever missed its binding site in the course of an 1D sliding round, in which case we would see a protein appear on one side of the site and translocate across it without binding enduringly. Owing to the redundancy of random walks, we would be able to determine only that the protein failed to fold on its target site a large ($\gtrsim 100 - 1000$) number of times rather than merely once. Measuring the time between association and full-site binding, knowing the location

both of the initial binding to DNA and the full-site, however, would allow us to infer roughly how many passes the protein had to make over the full-site before binding.

At the time, we also believed that our experiment required stronger or more clustered binding sites to yield a reliable slow-down signal. p53's affinity for the "accidental" full-sites on λ we estimated to be $4{\sim}5k_BT$ less than to the strongest-known biological binding site, the 5' binding site in the p21 promoter [141]. We made substantial progress toward realizing this project, described in the *Appendix* to this chapter, but upon re-analysis of the data used in Chapter 5, pursued that project instead.

6.1.2 More efficient data collection: fluctuations and multiplexing

Efficient data collection was hampered chiefly by two factors, the fluctuations in the DNA and the limited throughput of the assay. Fluctuations in the DNA were of an order of magnitude comparable to protein diffusion at short ($< 100ms$) timescales. The fluctuations introduced error in both our estimates of D and also in our assignment of particles to positions on DNA. They required us to increase our segment size, which entails greater averaging within a segment of features (*i.e.* half- and full-sites) and thus less dynamic range in the diffusion coefficients we can expect to observe.

Our group has had some success in creating doubly-tethered λ-DNA constructs by biotinylating both ends of the DNA. Sufficiently stretched double-tethered constructs have the advantage of lower-amplitude fluctuations, and would allow us to measure the diffusion of p53 on DNA in the absence of flow. They would increase our throughput on the one hand by not requiring us to exclude data from the 1/3 of the DNA farthest from the tether, but would decrease it on the other hand since the degree to which each DNA molecule is stretched would be distribution and so would require more complicated data analysis to map the positions of particles in the microscope to the contour of the DNA.

A throughput-increasing technique that lacks these disadvantages would be the use of "DNA curtains" recently developed by Fazio *et al.* [77]. In this technique, a microscopic linear berm is deposited on a coverslip, which is then coated with a lipid bilayer. The berm creates a break in the bilayer and allows functional groups to bind all in a line. Applying this technique to the research in Chapter 5 would have allowed me to greatly increase the DNA concentration, since all of the DNA would be aligned and thus there would be no need to distinguish individual DNA molecules. It would also speed the data analysis somewhat by allowing the omission of the step in which the proteins are assigned to DNA strands.

6.1.3 Fluorescence anisotropy

Although the work in Chapter 4 demonstrated that p53 does not spend enough time using the hopping mechanism for its D_{1D} to depend on ionic strength, it is nonetheless possible that it could hop infrequently. Another eukaryotic DNA-binding protein, PCNA, has been shown to translocate on DNA through a mixture of the two modes [142], and earlier theoretical work demonstrated that a mix of hopping and sliding could yield even greater acceleration of target site localization [143]. If a protein samples on average $\bar{n}$ base-pairs per sliding round, the redundancy can be reduced by allowing mesoscopic steps along with 1-bp steps. Occasional hops might also be relevant *in vivo* by allowing p53 or other proteins to translocate around roadblocks such and nucleosomes or other transcription factors. Indeed, DNA in p53's native environment is quite unlike the naked DNA used in our *in vitro* experiments.

The single-molecule microscopic technique we used lacks the resolution to visualize a protein momentarily leaving the DNA but rebinding locally. To measure the microscopic on- and off-rates of the protein from DNA, one technique available is fluorescence anisotropy. By exciting fluorophores of a suitable fluorescence lifetime with polarized light and measur-

ing the polarity of the light emitted, one distinguish between freely tumbling fluorophores, which will emit light the polarization of which will be randomized, and fluorophores with a fixed orientation, whose emisssions will retain some of the polarization information of the absorbed light. This technique can be implemented on a single-molecule level [144] to measure the distribution of on- and off-times of the protein from DNA, and thereby assess whether the protein indeed makes infrequent microscopic hops.

Single-molecule fluorescence anistropy would also allow the *in vitro* study of more-native DNA conformations. In most TIRFM experiments, including the ones presented in this thesis, it is necessary to stretch out DNA into a linear conformation. Our group [38] and others [19, 145] have examined the role of DNA conformation in protein-DNA diffusive search. Single-molecule data on the distribution of on- and off-times in compacted DNA could test and/or parametrize these models.

6.2 The need for *in vivo* and *in vivo*-like experiments

Some of the questions and techniques mentioned in the previous section (6.1) are motivated by the need to study systems in or closer to their native, *in vivo* state. The *in vivo* state of p53, for instance, is modulated by post-translational modifications, and, in instances of disease, by mutations. Beyond the state of the protein itself, the environment within a cell—particularly a eukaryotic cell—changes the nature of the search-and-recognition problem.

6.2.1 *In vivo* proteins: modifications and mutations

In vitro experiments on site-specific DBPs, including ours on p53, generally work with proteins lacking post-translational modifications. These modifications, however, may substantially alter the physical interaction of the proteins and DNA and thus the biological

consequences. Activation of p53, for example, requires phosphorylation in the N-terminal domain [84, 85], and in most cases acetylation of the C-terminal domain as well [146]. The N-terminal phosphorylations disrupt an interaction between p53 and a negative regulatory protein mdm2, while C-terminal acetylation directly affects its DNA-binding properties. *In vitro*, it increases its specific affinity [147] while reducing its non-specific affinity [148].

The measurements of binding lifetime, diffusivity, and number of sites scanned on DNA by p53 performed by our group, and their implications for p53's search-and-recognition mechanism, therefore, may not accurately describe p53's behavior *in vivo*, as will be discussed below. The effects of acetylation could be studied by repeating our experiments on p53, but first chemically or enzymatically acetylating it. A more precise and perhaps easier strategy would be mutating the protein, replacing one or more C-terminal lysines with a neutral amino acid, presumably alanine, or, as a steric mimic, glutamine.

Most known cancer mutations of p53 affect the core domain, and studies have focused on the effect of the mutations on the protein's binding to response elements. For example, a common mutation in a DNA-contacting moiety of the protein, R273H, decreases binding to the *gadd45* RE by three orders of magnitude [149]. The same mutation decreases non-specific affinity, however, only by a factor of 3–5. In the terms of the two-mode model, the mutation shifts the **R**/**S** equilibrium, $K_{R/S}$, in favor of the **S** mode. The model makes the testable prediction that this change should result in a larger effective diffusion coefficient, as the protein will spend less time in immobile in the **R** mode relative to the **S** mode.

In addition to the comparatively well-studied effects of core-domain mutations, it is possible that mutations in the C-terminal domain could modulate the protein's sliding behavior and thus its ability to bind its target sites in time to prevent tumorigenesis. Mutations that affect the rate of transition from the **S** mode to the **R** mode could also have clinically relevant consequences.

6.2.2 *In vivo* environments: Chromatin and other obstacles

The role of DNA conformation and crowding is particularly important for systems and molecules such as p53 that are of eukaryotic origin, owing to the highly chromatinized nature of eukaryotic DNA, and to the combinatorial nature of many biological processes involving DNA, such as transcriptional activation. While the chromatinization of eukaryotic DNA excludes a large (95%~99%) fraction of DNA from search by site-specific DBPs, it also creates obstacles in the way of diffusing proteins. The presence of nucleosomes extrinsically limits $\bar{n}$, unless a protein can hop around them.

Owing to experimental difficulty, few *in vivo* single-molecule imaging experiments of proteins diffusing on DNA have been conducted. An intermediate regime between the *in vitro* experiments that have dominated the field and imaging live cells would be to reconstitute nucleosomes in *in vitro* assays, and observe transcription-factor or damage-repair proteins' dynamics in their presence. If an experimental setup similar to ours is employed, but with nucleosomes labeled with a dye of a different color, the behavior of nucleosomes and p53 or other DBPs could be studied when they encounter each other.

In addition to serving as roadblocks *near* a target site, nucleosomes may directly cover one or more target sites. It has been suggested that nucleosomes could function to decrease the variance in transcriptional activation levels by competing with transcription factors for the TFs' binding sites [150]. This would suppress activation at low TF concentrations. Nucleosomes might also function as an evolutionarily inexpensive way to achieve combinatorial gene regulation—if a nuclesome covers the binding sites of more than one TF, then even if the TFs have no evolved interaction, they will effectively interact by "teaming up" to displace the nucleosome. Single-molecule microscopic studies of TFs competing with nucleosomes on DNA could elucidate the mechanism of this competition. For example, it has been proposed that multiple transcription factors might passively displace a nucle-

osome by the first transcription factor binding to a site that is exposed by a nucleosome transiently unwinding, which then favors further unwinding and further binding by the next transcription factor [151]. This sort of mechanism, where multiple TFs cause a nuclesome to roll off of their target sites, would require the TFs to approach the nuclesome from the same side. Assessing whether displacement requires TFs on the same side would be difficult using bulk biological techniques, and would likely involve perturbing the DNA in some way, while the direction of approach by TFs toward a nucleosome can be easily determined using single-molecule microscopy.

Another consequence of crowded DNA is that multiple diffusing species may serve as mobile reflecting barriers, substantially altering each other's sliding kinetics. It has been suggested [152] that the effect of such "jamming" is to cause proteins' random walks to become subdiffusive on the mesoscale ($\gtrsim$ the inter-protein distance). *I.e.,* the number of base-pairs visited would not go as the square-root of time (Equation 1.4), but rather $\propto \tau_{1D}^{\gamma}$, with $\gamma < \frac{1}{2}$. Single-molecule experiments very similar to ours could be conducted with a higher protein concentration, such that proteins are likely to encounter others on DNA several times within τ_{1D}, and plots of particles' MSDs versus Δt examined for subdiffusive behavior as a function of protein concentration. This would likely necessitate a low labeling efficiency of the proteins in order to reduce background and to be able to identify individual particles on DNA. The non-observation of subdiffusivity despite protein concentrations high enough that jamming should be frequent would imply that the proteins hop sufficiently frequently as to circumvent obstacles.

It is also worth noting that most *in vitro* experiments examining DBPs undergoing 1D random walks on DNA use lower salt concentrations than the 150–200 mM that is found *in vivo* [153]. Our work on the aggregate sliding properties of p53 (Chapter 4), for example, used an ionic strength of 125 mM, in order to obtain longer trajectories. Subsequent

work [53] on a p53 construct consisting of the tetramerization and C-terminal domains could not be conducted at salt concentrations greater than 75 mM owing to prohibitively short lifetimes on DNA, and attempts to visualize the C-terminal peptide alone succeeded only at 13 mM ionic strength.

Lifetimes on DNA, τ_{1D}, may, however, in nature be so short that physiologically appropriate salt concentrations are challenging with standard single-molecule microscopic techniques. In a study of the search process of *lac* repressor in *E. coli*, a 1D lifetime of $10^{-4} - 10^{-2}$s was estimated [42], the shorter end of which is at the limit of current EM-CCD technology. Reported in the past year, a similar *in vivo* experiment [52] on the yeast transcription factor Mbp1, however, reported mean lifetime of 0.8s, and the mean lifetime of p53 on DNA observed in the work in Chapter 5 at 163 mM salt concentration is ~0.9s, although the acetylation of p53's C-terminal lysines, which is necessary for activation *in vivo*, is known to decrease its lifetime on DNA [148].

6.3 The two-mode model and eukaryotes

The relatively large values of τ_{1D} observed for eukaryotic transcription factors [53, 114, 52], combined with the possibility of $\bar{n}$ being extrinsically reduced by obstacles on crowded eukaryotic DNA, suggests that some eukaryotic transcription factors may not need kinetic pre-selection to find their target sites efficiently (section 2.4). That is, they may need in addition to a highly sequence-dependent **R** landscape nothing more than a flat or uncorrelated **S** energy landscape (section 2.3.2).

Experimental values from our group's *in vitro* studies of p53 truncation mutants suggest that the protein need not transition from the **S** to the **R** mode any faster than at a rate of 700s^{-1}, even without pre-selection, which is within the range of experimentally

estimated k_f's for non-specific-to-specific conformational changes by a transcription factor, 10^3~10^5 s^{-1} [55, 54]. As mentioned above, however, physiologically active p53's lifetime on DNA is lower, because of the reduction of electrostatic interactions with DNA owing to the acetylation of C-terminal lysine residues required for the TF's activation, and so the *in vivo* requirements for efficient folding are almost certainly more demanding. Results from *in vivo* studies on a yeast transcription factor, by Larson *et al.*, similarly give a sufficiently low minimum folding rate, 3×10^{-4}s, for efficient search.

6.4 Disordered proteins and accelerated binding to DNA

The acceleration of TF–cognate-site binding by conformational flexibility in the protein illustrates the importance of the coupling of folding and binding for molecular search and recognition. A plurality of the attention that accelerated protein-DNA binding owing to protein conformation flexibility has received recently has considered the acceleration in terms of the "fly-casting" model, by which a partially unfolded protein has an expanded capture radius [154] and is guided electrostatically upon approach toward the folded, binding conformation. A critical analysis by Huang *et al.* of this mechanism, however, has argued that increasing the capture radius for disordered proteins requires smaller D_{3D}, counteracting the acceleration of target-binding [155]. Huang *et al.* argue that rather than an increased capture radius, the source of acceleration is that fewer encounters between protein and DNA are required; a flexible conformation allows the protein to orient properly on the approach. Referring to the expression for the diffusion-limited rate of molecular association (Equation 1.1, reprinted here for convenience), the conformation flexibility increases a rather than b, where b is the linear size of the target, and a is the fraction of collisions

resulting in binding:

$$k_{\text{smol}} = 4\pi D_{3D} b a \tag{6.1}$$

While one function of non-specific, **S**-mode binding is that it increases the capture radius from 1 base-pair to 10^2~10^3 bp, it also serves to increase a by relying on an orientation-insensitive mechanism of binding. For purely-3D association, an attempt at recognition is successful only if the protein comes within b (0.34 nm) of the target site, *and* if the protein has the correct orientation. Sequence-specific binding of proteins to DNA depends at least in part on some combination of hydrogen-bond interactions, pi-interactions, and hydrophobic-surface interactions, all of which are highly sensitive to orientation. If a protein can bind non-specifically to DNA through electrostatic interactions, however, then successful binding is less sensitive to its orientation upon approach.

Within the past year, increased flexibility has been shown directly to accelerate the binding of a transcriptional regulator to DNA. The human papilloma virus (HPV) E2 protein forms a non-specific complex based on electrostatic interactions upon encountering DNA, and then folds into specific conformation on its target site. Brown *et al.* compared wild-type E2 to a mutant in which they deleted two leucine residues that form part of the protein's hydrophobic core. They found using ^{15}N NMR relaxation and hydrogen/deuterium exchange that this mutation had the effect in solution of destabilizing the core as well as the anchor points of the unstructured, positively-charged loop regions responsible for non-specific DNA binding, and with stop-flow kinetics experiments that the mutant associated approximately 6 times faster to DNA than did the wild type. The specific mutant-DNA complex, however, was structurally similar to the specific wild-type–DNA complex.

As the C-terminal domain of p53 is similarly unstructured in solution and interacts with DNA through its positively-charged arginine and lysine residues, it may be conjectured that p53 binds DNA in much the same was as HPV E2—an initial binding to non-specific

DNA via electrostatic contacts in a disordered domain (or region) followed by specific binding with a structured domain. Indeed, the C-terminal domain has been shown to fold upon binding to S100B($\beta\beta$) [156], an inhibitor of protein kinase C [157], which phosphorylates the C-terminus. If it behaves similarly upon binding to DNA, then p53's association to its specific site can be described as two sequential mechanisms of folding coupled to binding: the first upon binding from solution to DNA non-specifically, and the second upon reaching and recognizing its target site and folding from the **S** conformation to the **R** conformation. Interestingly, if the C-terminus becomes ordered upon binding non-specifically, it nonetheless loses its order when the protein is bound specifically at a target site, as revealed by EM structures of the protein on DNA [47].

That p53 and HPV E2 might bind and recognize DNA by the same mechanism is both deep and ironic. E2 is a viral protein, while p53 a high-eukaryotic one. A common mechanism of target-site binding, along with evidence for such a mechanism in bacteria and unicellular eukaryotes, suggests that a multi-mode model as discussed in Part I can serve as a description of protein-DNA interactions fundamental to life generally. The irony is that while p53 and E2 may reach and recognize their target sites in similar ways, the consequences of their successful binding and transcriptional activation are antithetical. Among other functions, p53 induces apoptosis of cells in danger of becoming oncogenic. The cell thus dies so that the organism might flourish. E2, however, induces the transcription of its fellow HPV proteins E6 and E7 [158]. The former interacts with a host ubiquitin ligase to target p53 for degradation, and the latter inactivates p53's partner in tumor suppression, retinoblastoma protein, which halts the cell cycle in the G1 phase in response to DNA damage. Facilitated target-site search and recognition by means of multiple protein-DNA modes of interaction is thus likely instrumental in both tumor suppression and oncogenesis.

6.5 Acknowledgements

Anna Kochaniak was indispensible in providing advice, instruction, and reagents for large components of the project outlined in the *Appendix*. Dr. Candice Etson was similarly helpful in troubleshooting and providing help with assays and other techniques. Aid with biochemical and molecular-biological techniques and helpful converations were provided by the aforementioned persons as well as by Irene Kim, Dr. Nathan Tanner, Dr. Joseph Loparo, Dr. Mark Elenko, Dr. Daniel Floyd, Dr. Satoshi Habuchi, and Dr. Samir Hamdan.

6.A Appendix

This appendix describes progress toward the synthesis of and experiments using a DNA construct bearing the p21 5' response element, the strongest functional p53 RE known in nature.

6.A1 DNA construct

The desired DNA construct would have a length on the order of λ phage's genome (16 μm, 48 kb) so that it could be stretched adequately by flow, and have multiple but adequately spaced potential wells for p53. I intended to make the construct by annealing a biotinylated oligo to circular DNA bearing the potential well, and extend the DNA using rolling-circle amplification (RCA) with the T7 phage DNA replisome. I attempted three main strategies for synthesizing this construct.

Biochemical insertion of p21 site into m13 vector. Our lab had previously devised an RCA assay for measuring replication rates using a biotinylated DNA tail annealed to a commerically available circular single-stranded DNAvector (m13). My strategy was to use T7 polymerase to extend the tail and thereby make the DNA double stranded, use restriction enzymes to cut out part of the vector and leave sticky ends for an insert bearing the p53 site, ligate the insert into the vector, and then subject the vector to RCA (Figure 6.1). This strategy was not successful, owing to extremely inefficient ligation.

Attachment of biotinylated tail to gifted p21 vector. My next attempt was to acquire a plasmid bearing the desired p53 site, and attach a biotinylated tail to it. I would treat the plasmid with a set of nicking endonucleases that produced nearby nicks and treat with an excess of the tail, and then ligate the tail to the plasmid (Figure 6.2). This strategy

suffered from inefficient displacement of the DNA between the nearby nicks by the tail, most likely due to poor annealing by the tail, as well as inefficient loading by the T7 replisome to extend the tail. I tried to clear out more room for the tail but using a nicking endonuclease that had a single site on the plasmid and then treat with an exonuclease that loaded on nicks such as the Klenow fragment or DNA ExoIII, as well as E. coli PolI to clear out room for the replisome to load. None of these strategies (and others) produced efficient RCA products.

Custom-designed vector with cloning. I was finally able to construct the RCA product by cloning into a vector an insert containing the desired p53 site as well as two optimally spaced nicking sites (Figure 6.3).

Repair of rolling-circle DNA Rolling-circle amplification leaves RNA primers and discontinuities on the lagging strand. These could interfere with p53's diffusion, so I needed to repair my amplified DNA. To this end, I treated the RCA product with *E. coli* DNA PolI, which has a 5'→3' exonuclease and leaves no gap between newly incorporated deoxyribonucleotides and DNA in front of it. I assessed primer replacement by using radiolabeled nucleotides and measuring the extent of radioincorporation.

After replacing RNA primers with DNA, I repaired the nicks in the RCA product with T4 ligase. One way to assess the ligation is with pulsed-field gel electrophoresis of DNA treated with S1 or micrococcal nuclease, which are endonucleases that cut nicked DNA but have very limited activity on dsDNA. Since the purpose of ligation is to make sure that nicks do not interfere with p53 sliding, ligation may not be necessary, however, if I can observe p53 sliding over nicks. To see whether it could, I conducted a single-molecule microscopy experiment of p53 on a control PolI- and ligase-treated RCA product that had a scrambled p53 binding site. If the p53 slid freely, then either ligation was successful or

nicks did not impede p53's sliding. At this point, however, I discovered that the labeled p53 I had been using had degraded and no longer slid, even on unnicked control DNA (λ phage).

6.A2 Binding of p53 to target DNA

Even if my protein no longer slid, I verified that it indeed bound the target site in my flow-cell setup by observing binding of the protein to RCA product with a periodicity equal to the length of the vector, that is, the spacing between binding sites. At this point, the project was suspended in favor of improving the data analysis used in the work discussed in Chapter 5.

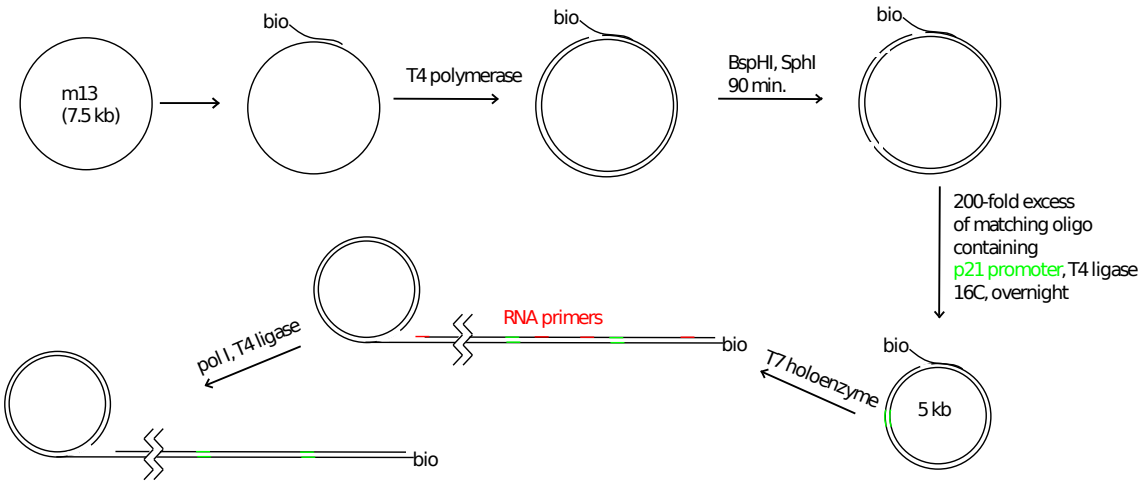


Figure 6.1: Scheme for making DNA construct by inserting p53 binding site using commercial enzymes. I started with an m13 genome and annealed to it a biotinylated 33-bp oligonucleotide. The construct was made double-stranded using T7 polymerase, and then cleaved with restriction endonucleases BspHI and SphI. A 200-fold excess of an oligonucleotide with sticky ends matching those left by BspHI and SphI was ligated with T4 ligase overnight. The ligation proved prohibitively inefficient, however.

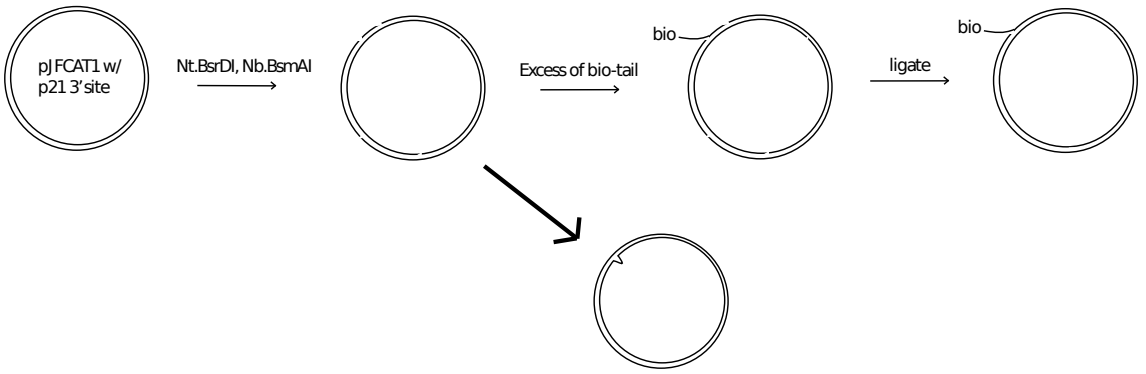


Figure 6.2: Scheme for making DNA construct using a plasmid containing the p21 3' response element, obtained as a gift from the laboratory of Dr. Wafik El-Deiry. The plasmid was nicked with endonucleases Nt.BsrDI and Nb.BsmAI, which cut 6bp away from each other on the same strand. The attempted product of the next step, annealing with a biotinylated tail and ligation with T4 ligase, was not succesful as the major product of the reaction was an intra-plasmid ligation product, identified by gel electrophoresis as a ladder of linking-number isomers.

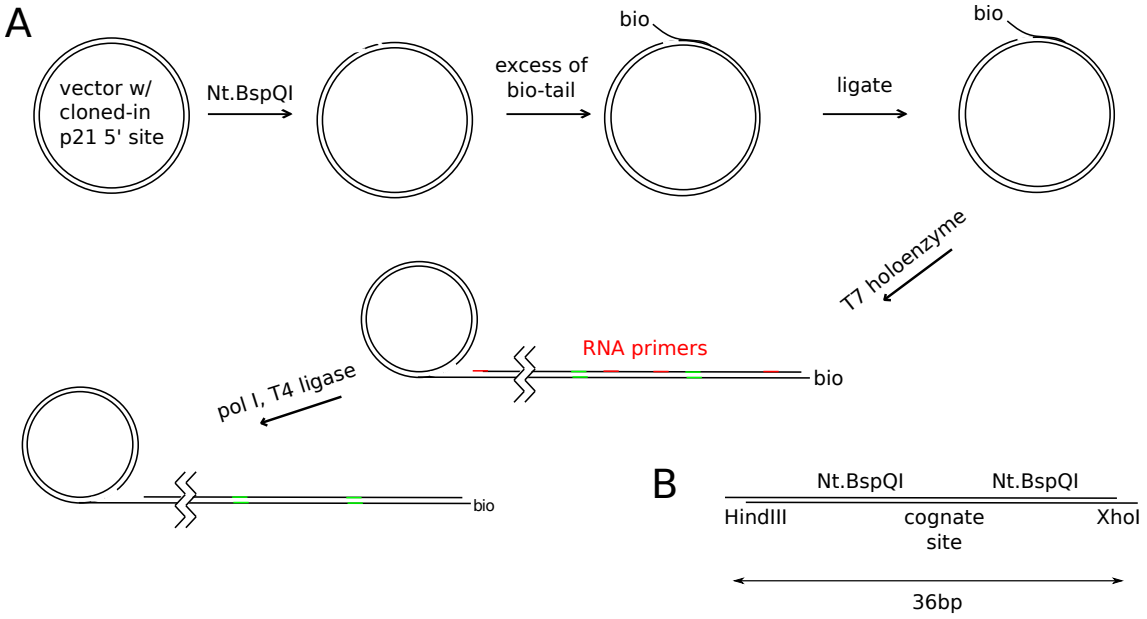


Figure 6.3: (**A**): Scheme for making DNA construct using cloning. I designed an oligonucleotide (**B**) to clone into the vector that included two rare (7-letter) nicking sites and the p21 5' RE. After cloning, I treated the construct with nicking endonuclease Nt.BspQI, producing nicks 26 bp apart, and annealed an excess of a biotinylated tail oligo. This strategy was successful, and I was able to use rolling-circle amplification to extend the construct.

Bibliography

[1] Cavasotto CN, W Orry AJ (May) Ligand docking and structure-based virtual screening in drug discovery. *Curr. Topics Med. Chem.* 7:1006–1014.

[2] B-Rao C, Subramanian J, Sharma SD (2009) Managing protein flexibility in docking and its applications. *Drug Discovery Today* 14:394 – 400.

[3] Shoichet BK, McGovern SL, Wei B, Irwin JJ (2002) Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* 6:439 – 446.

[4] Alvarez JC (2004) High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* 8:365 – 370.

[5] Jones S, Thornton JM (1995) Protein-protein interactions: A review of protein dimer structures. *Progress in Biophysics and Molecular Biology* 63:31 – 65.

[6] Jones S, Thornton JM (1996) Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* 93:13–20.

[7] Shoemaker BA, Panchenko AR (2007) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comp. Biol.* 3:e42.

[8] Shoemaker BA, Panchenko AR (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comp. Biol.* 3:e43.

[9] Jones S, van Heyningen P, Berman HM, Thornton JM (1999) Protein-DNA interactions: a structural analysis. *J. Mol. Biol.* 287:877 – 896.

[10] Stormo GD, Zhao Y (2010) Determining the specificity of protein-DNA interactions. *Nature reviews. Genetics* 11:751–760.

[11] Rohs R, et al. (2010) Origins of specificity in protein-DNA recognition. *Annual Review of Biochemistry* 79:233–269.

[12] Riggs AD, Bourgeois S, Cohn M (1970) The Lac repressor-operator interaction. 3. Kinetic studies. *J. Mol. Biol.* 53:401–417.

[13] Adam, G. & Delbrück M (1968) in *Structural Chemistry and Molecular Biology*, ed Rich, A. and Davidson, N. (W.H. Freeman and Company), pp 198–215.

[14] Riggs AD, Bourgeois S, Cohn M (1970) The lac repressor-operator interaction. 3. kinetic studies. *J Mol Biol* 53:401–417.

[15] Richter PH, Eigen M (1974) Diffusion controlled reaction rates in spheroidal geometry. application to repressor–operator association and membrane bound enzymes. *Biophys Chem* 2:255–263.

[16] Berg OG, Blomberg C (1976) Association kinetics with coupled diffusional flows. special application to the lac repressor–operator system. *Biophys Chem* 4:367–381.

[17] Berg OG, Winter RB, von Hippel PH (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* 20:6929–6948.

[18] Revzin A (1990) *The Biology of Nonspecific DNA Protein Interactions* (CRC-Press).

[19] Hu T, Grosberg AY, Shklovskii BI (2006) How proteins search for their specific sites on DNA: the role of DNA conformation. *Biophys. J.* 90:2731–2744.

[20] Jack WE, Terry BJ, Modrich P (1982) Involvement of outside DNA sequences in the major kinetic path by which EcoRI endonuclease locates and leaves its recognition sequence. *Proc. Natl. Acad. Sci. USA* 79:4010–4014.

[21] Gowers DM, Wilson GG, Halford SE (2005) Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA. *Proc. Natl. Acad. Sci. USA* 102:15883–15888.

[22] Wang YM, Austin RH, Cox EC (2006) Single molecule measurements of repressor protein 1D diffusion on DNA. *Phys Rev Lett* 97:048302.

[23] Gorman J, Greene EC (2008) Visualizing one-dimensional diffusion of proteins along dna. *Nat Struct Mol Biol* 15:768–774.

[24] Granli A, Yeykal CC, Robertson RB, Greene EC (2006) Long-distance lateral diffusion of human rad51 on double-stranded dna. *Proc Natl Acad Sci U S A* 103:1221–1226.

[25] Blainey PC, van Oijen AM, Banerjee A, Verdine GL, Xie XS (2006) A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proc Natl Acad Sci U S A* 103:5752–5757.

[26] de Saro FJL, Marinus MG, Modrich P, O'Donnell M (2006) The beta sliding clamp binds to multiple sites within MutL and MutS. *J. Biol. Chem.* 281:14340–14349.

[27] Slutsky M, Mirny LA (2004) Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys J* 87:4021–4035.

[28] Halford SE, Marko JF (2004) How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res* 32:3040–3052.

[29] Coppey M, Bnichou O, Voituriez R, Moreau M (2004) Kinetics of target site localization of a protein on DNA: a stochastic approach. *Biophys J* 87:1640–1649.

[30] Hu T, Shklovskii BI (2006) How does a protein search for the specific site on DNA: The role of disorder. *Phys Rev E Stat Nonlin Soft Matter Phys* 74:021903.

[31] Hu T, Grosberg AY, Shklovskii BI (2006) How proteins search for their specific sites on DNA: the role of DNA conformation. *Biophys J* 90:2731–2744.

[32] Lomholt MA, Ambjrnsson T, Metzler R (2005) Optimal target search on a fast-folding polymer chain with volume exchange. *Phys Rev Lett* 95:260603.

[33] Lomholt MA, van den Broek B, Kalisch SMJ, Wuite GJL, Metzler R (2009) Facilitated diffusion with DNA coiling. *Proc Natl Acad Sci U S A* 106:8204–8208.

[34] Zhou HX, Szabo A (2004) Enhancement of association rates by nonspecific binding to DNA and cell membranes. *Phys Rev Lett* 93:178101.

[35] Loverdo C, et al. (2009) Quantifying hopping and jumping in facilitated diffusion of DNA-binding proteins. *Phys Rev Lett* 102:188101.

[36] Bonnet I, et al. (2008) Sliding and jumping of single EcoRV restriction enzymes on non-cognate DNA. *Nucleic Acids Res* 36:4118–4127.

[37] Mirny L, et al. (2009) How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. A* 42:434013.

[38] Wunderlich Z, Mirny LA (2008) Spatial effects on the speed and reliability of protein-DNA search. *Nucleic Acids Res* 36:3570–3578.

[39] Slutsky M, Mirny LA (2004) Kinetics of protein-DNA interaction: Facilitated target location in sequence-dependent potential. *Biophys. J.* 87:4021–4035.

[40] Slutsky M, Kardar M, Mirny LA (2004) Diffusion in correlated random potentials, with applications to DNA. *Phys Rev E Stat Nonlin Soft Matter Phys* 69:061903.

[41] Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* 193:723–750.

[42] Elf J, Li GW, Xie XS (2007) Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* 316:1191–1194.

[43] Tafvizi A, et al. (2008) Tumor suppressor p53 slides on DNA with low friction and high stability. *Biophys. J.* 95:L01–3.

[44] Iwahara J, Zweckstetter M, Clore GM (2006) NMR structural and kinetic characterization of a homeodomain diffusing and hopping on nonspecific DNA. *Proceedings of the National Academy of Sciences* 103:15062–15067.

[45] Kalodimos C, et al. (2004) Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* 305:386–389.

[46] Banerjee A, Yang W, Karplus M, Verdine GL (2005) Structure of a repair enzyme interrogating undamaged DNA elucidates recognition of damaged DNA. *Nature* 434:612–618.

[47] Melero R, et al. (2011) Electron microscopy studies on the quaternary structure of p53 reveal different binding modes for p53 tetramers in complex with DNA. *Proceedings of the National Academy of Sciences* 108:557–562.

[48] Kalodimos C, Boelens R, Kaptein R (2004) Toward an integrated model of protein-DNA recognition as inferred from NMR studies on the lac repressor system. *Chem. Rev.* 104:3567–3586.

[49] Viadiu H, Aggarwal AK (2000) Structure of bamhi bound to nonspecific DNA: a model for DNA sliding. *Mol Cell* 5:889–895.

[50] Townson SA, Samuelson JC, Bao Y, yong Xu S, Aggarwal AK (2007) Bstyi bound to noncognate DNA reveals a "hemispecific" complex: Implications for DNA scanning. *Structure* 15:449–459.

[51] Erie D, Yang G, Schultz H, Bustamante C (1994) DNA bending by Cro protein in specific and nonspecific complexes: implications for protein site recognition and specificity. *Science* 266:1562–6.

[52] Larson DR, Zenklusen D, Wu B, Chao JA, Singer RH (2011) Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science* 332:475–478.

[53] Tafvizi A, Huang F, Fersht AR, Mirny LA, van Oijen AM (2010) A single-molecule characterization of p53 search on DNA. *Proceedings of the National Academy of Sciences* pp 563–568.

[54] Salinas RK, Diercks T, Kaptein R, Boelens R (2006) Cooperative alpha-helix unfolding in a protein-DNA complex from hydrogen-deuterium exchange. *Protein Sci* 15:1752–1759.

[55] Kalodimos CG, Boelens R, Kaptein R (2002) A residue-specific view of the association and dissociation pathway in protein–DNA recognition. *Nat Struct Biol* 9:193–197.

[56] Akke M (2002) NMR methods for characterizing microsecond to millisecond dynamics in recognition and catalysis. *Curr. Opin. Struct. Biol.* 12:642–647.

[57] Spolar RS, Record MT (1994) Coupling of local folding to site-specific binding of proteins to DNA. *Science* 263:777–784.

[58] Qi Y, et al. (2009) Encounter and extrusion of an intrahelical lesion by a DNA repair enzyme. *Nature* 462:762–766.

[59] Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.* 1:1–37.

[60] Stivers JT, Pankiewicz KW, Watanabe KA (1999) Kinetic mechanism of damage site recognition and uracil flipping by escherichia coli uracil DNA glycosylase. *Biochemistry* 38:952–963.

[61] Blainey PC, van Oijen AM, Banerjee A, Verdine GL, Xie XS (2006) A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proc. Natl. Acad. Sci. USA* 103:5752–5757.

[62] Schurr JM (1979) The one-dimensional diffusion coefficient of proteins absorbed on DNA. hydrodynamic considerations. *Biophys Chem* 9:413–414.

[63] Kim JG, Takeda Y, Matthews BW, Anderson WF (1987) Kinetic studies on Cro repressor-operator DNA interaction. *J. Mol. Biol.* 196:149–158.

[64] Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.* 22:403–434.

[65] Kuznetsov V, Orlov Y, Wei C, Ruan Y (2007) Computational analysis and modeling of genome-scale avidity distribution of transcription factor binding sites in ChIP-PET experiments. *Genome Inform.* 19:83–94.

[66] Hopfield JJ (1974) Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. USA* 71:4135–4139.

[67] Kolesov G, Wunderlich Z, Laikova ON, Gelfand MS, Mirny LA (2007) How gene order is influenced by the biophysics of transcription regulation. *Proc. Natl. Acad. Sci. USA* 104:13948–13953.

[68] McKinney K, Mattia M, Gottifredi V, Prives C (2004) p53 linear diffusion along DNA requires its C terminus. *Mol. Cell* 16:413–424.

[69] Y H, et al. (1999) Single-molecule imaging of RNA polymerase-DNA interactions in real time. *Biophys. J.* 76:709–715.

[70] Kim JH, Larson RG (2007) Single-molecule analysis of 1D diffusion and transcription elongation of T7 RNA polymerase along individual stretched DNA molecules. *Nucleic Acids Res.* 35:3848–3858.

[71] Winter RB, Berg OG, von Hippel PH (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor–operator interaction: kinetic measurements and conclusions. *Biochemistry* 20:6961–6977.

[72] Halford SE, Szczelkun MD (2002) How to get from A to B: strategies for analysing protein motion on DNA. *Eur Biophys J* 31:257–267.

[73] Kochaniak AB, et al. (2009) Proliferating cell nuclear antigen (PCNA) uses two dstinct modes to move along the DNA. *J. Biol. Chem.*

[74] Elf J, Li GW, Xie XS (2007) Probing transcription factor dynamics at the single-molecule level in a living cell. *Science* 316:1191–1194.

[75] Wang YM, Austin RH, Cox EC (2006) Single molecule measurements of repressor protein 1D diffusion on DNA. *Phys Rev Lett* 97:048302.

[76] Iwahara J, Zweckstetter M, Clore GM (2006) NMR structural and kinetic characterization of a homeodomain diffusing and hopping on nonspecific DNA. *Proc Natl Acad Sci U S A* 103:15062–15067.

[77] Fazio T, Visnapuu ML, Wind S, Greene EC (2008) DNA curtains and nanoscale curtain rods: high-throughput tools for single molecule imaging. *Langmuir* 24:10524–10531.

[78] Thompson RE, Larson DR, Webb WW (2002) Precise nanometer localization analysis for individual fluorescent probes. *Biophys J* 82:2775–2783.

[79] Underhill PT, Doyle PS (2004) On the coarse-graining of polymers into bead-spring chains. *J. Non-Newtonian Fluid Mech.* 122:3–31 XIIIth International Workshop on Numerical Methods for Non-Newtonian Flows.

[80] Russo MT, et al. (2004) Accumulation of the oxidative base lesion 8-hydroxyguanine in DNA of tumor-prone mice defective in both the Myh and Ogg1 DNA glycosylases. *Cancer Res.* 64:4411–4414.

[81] Stiewe T (2007) The p53 family in differentiation and tumorigenesis. *Nat. Rev. Cancer* 7:165–167.

[82] Vousden K, Lane D (2007) p53 in health and disease. *Nat. Rev. Mol. Cell. Biol.* 8:275–283.

[83] Vogelstein B, Lane D, Levine AJ (2000) Surfing the p53 network. *Nature* 408:307–310.

[84] Lees-Miller SP, Sakaguchi K, Ullrich SJ, Appella E, Anderson CW (1992) Human dna-activated protein kinase phosphorylates serines 15 and 37 in the amino-terminal transactivation domain of human p53. *Molecular and Cellular Biology* 12:5041–5049.

[85] Baudier J, Delphin C, Grunwald D, Khochbin S, Lawrence JJ (1992) Characterization of the tumor suppressor protein p53 as a protein kinase C substrate and a S100b-binding protein. *Proceedings of the National Academy of Sciences* 89:11627–11631.

[86] Wang Y, Schwedes JF, Parks D, Mann K, Tegtmeyer P (1995) Interaction of p53 with its consensus DNA-binding site. *Mol Cell Biol* 15:2157–2165.

[87] Jayaraman L, Prives C (1999) Covalent and noncovalent modifiers of the p53 protein. *Cell Mol Life Sci* 55:76–87.

[88] Weinberg RL, Veprintsev DB, Fersht AR (2004) Cooperative binding of tetrameric p53 to DNA. *Journal of Molecular Biology* 341:1145–1159.

[89] Hupp TR, Meek DW, Midgley CA, Lane DP (1992) Regulation of the specific DNA-binding function of p53. *Cell* 71:875–886.

[90] McKinney K, Mattia M, Gottifredi V, Prives C (2004) p53 linear diffusion along dna requires its c terminus. *Mol Cell* 16:413–424.

[91] Klein C, et al. (2001) NMR spectroscopy reveals the solution dimerization interface of p53 core domains bound to their consensus DNA. *J. Biol. Chem.* 276:49020–49027.

[92] Menendez D, Inga A, Resnick MA (2009) The expanding universe of p53 targets. *Nat. Rev. Cancer* 9:724–737.

[93] Qian H, Wang T, Naumovski L, Lopez CD, Brachmann RK (2002) Groups of p53 target genes involved in specific p53 downstream effects cluster into different classes of DNA binding sites. *Oncogene* 21:7901–7911.

[94] Göhler T, et al. (2002) Specific interaction of p53 with target binding sites is determined by DNA conformation and is regulated by the C-terminal domain. *J. Biol. Chem.* 277:41192–41203.

[95] Wei CL, et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell* 124:207–219.

[96] Ma B, Pan Y, Zheng J, Levine AJ, Nussinov R (2007) Sequence analysis of p53 response-elements suggests multiple binding modes of the p53 tetramer to DNA targets. *Nucleic Acids Res.* 35:2986–3001.

[97] Horvath MM, Wang X, Resnick MA, Bell DA (2007) Divergent evolution of human p53 binding sites: Cell cycle versus apoptosis. *PLoS Genet.* 3:e127.

[98] Noureddine MA, et al. (2009) Probing the functional impact of sequence variation on p53-DNA interactions using a novel microsphere assay for protein-DNA binding with human cell extracts. *PLoS Genet.* 5:e1000462.

[99] Nikolova PV, Henckel J, Lane DP, Fersht AR (1998) Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. *Proc. Natl. Acad. Sci. USA* 95:14675–14680.

[100] Bullock AN, et al. (1997) Thermodynamic stability of wild-type and mutant p53 core domain. *Proc. Natl. Acad. Sci. USA* 94:14338–14342.

[101] Natan E, Hirschberg D, Morgner N, Robinson CV, Fersht AR (2009) Ultraslow oligomerization equilibria of p53 and its implications. *Proc. Natl. Acad. Sci. USA* 106:14327–14332.

[102] Zhou BB, Elledge SJ (2000) The DNA damage response: putting checkpoints in perspective. *Nature* 408:433–439.

[103] Weinberg RL, Freund SMV, Veprintsev DB, Bycroft M, Fersht AR (2004) Regulation of DNA binding of p53 by its C-terminal domain. *J Mol Biol* 342:801–811.

[104] McKinney K, Prives C (2002) Efficient specific dna binding by p53 requires both its central and c-terminal domains as revealed by studies with high-mobility group 1 protein. *Mol Cell Biol* 22:6797–6808.

[105] Sauer M, et al. (2008) C-terminal diversity within the p53 family accounts for differences in DNA binding and transcriptional activity. *Nucleic Acids Res* 36:1900–1912.

[106] Espinosa JM, Emerson BM (2001) Transcriptional regulation by p53 through intrinsic DNA/chromatin binding and site-directed cofactor recruitment. *Mol Cell* 8:57–69.

[107] Weinberg RL, Veprintsev DB, Fersht AR (2004) Cooperative binding of tetrameric p53 to DNA. *J Mol Biol* 341:1145–1159.

[108] Liu Y, Lagowski JP, Vanderbeek GE, Kulesz-Martin MF (2004) Facilitated search for specific genomic targets by p53 c-terminal basic dna binding domain. *Cancer Biol Ther* 3:1102–1108.

[109] Crook T, Marston NJ, Sara EA, Vousden KH (1994) Transcriptional activation by p53 correlates with suppression of growth but not transformation. *Cell* 79:817–827.

[110] Liu Y, Kulesz-Martin MF (2006) Sliding into home: facilitated p53 search for targets by the basic DNA binding domain. *Cell Death Differ* 13:881–884.

[111] Widom J (2005) Target site localization by site-specific, DNA-binding proteins. *Proc Natl Acad Sci U S A* 102:16909–16910.

[112] Kolesov G, Wunderlich Z, Laikova ON, Gelfand MS, Mirny LA (2007) How gene order is influenced by the biophysics of transcription regulation. *Proc Natl Acad Sci U S A* 104:13948–13953.

[113] Kleinhans D, Friedricha R (2007) Maximum likelihood estimation of drift and diffusion functions. *Physics Letters A* 368:194–198.

[114] Leith J, et al. (2012) Sequence-dependent sliding kinetics of p53. *Submitted to Proc. Natl. Acad. Sci. USA*.

[115] Schurr JM (1979) The one-dimensional diffusion coefficient of proteins absorbed on DNA. hydrodynamic considerations. *Biophys Chem* 9:413–414.

[116] van Oijen AM, et al. (2003) Single-molecule kinetics of lambda exonuclease reveal base dependence and dynamic disorder. *Science* 301:1235–1238.

[117] Lee JB, et al. (2006) DNA primase acts as a molecular brake in DNA replication. *Nature* 439:621–624.

[118] Nikolova PV, Henckel J, Lane DP, Fersht AR (1998) Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. *Proc Natl Acad Sci U S A* 95:14675–14680.

[119] Veprintsev DB, et al. (2006) Core domain interactions in full-length p53 in solution. *Proc Natl Acad Sci U S A* 103:2115–2119.

[120] Qian H, Sheetz MP, Elson EL (1991) Single particle tracking. analysis of diffusion and flow in two-dimensional systems. *Biophys J* 60:910–921.

[121] Creighton, T. E. FWH (1997) *Proteins Structures and Molecular Properties.* (Oxford University Press).

[122] Berg HC (1993) *Random walks in biology* (Princeton, Princeton University Press).

[123] Doyle PS, Ladoux B, Viovy JL (2000) Dynamics of a tethered polymer in shear flow. *Phys Rev Lett* 84:4769–4772.

[124] Smith SB, Finzi L, Bustamante C (1992) Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* 258:1122–1126.

[125] Berg OG, Winter RB, von Hippel PH (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* 20:6929–6948.

[126] von Hippel PH, Berg OG (1989) Facilitated target location in biological systems. *J. Biol. Chem.* 264:675–678.

[127] Liu Y, Lagowski J, Vanderbeek G, Kulesz-Martin M (2004) Facilitated search for specific genomic targets by p53 C-terminal basic DNA binding domain. *Cancer Biol. Ther.* 3:1102–1108.

[128] Winter RB, Berg OG, von Hippel PH (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor–operator interaction: kinetic measurements and conclusions. *Biochemistry* 20:6961–6977.

[129] Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315:233–237.

[130] Zhang C, et al. (2006) A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Res.* 34:2238–2246.

[131] Veprintsev DB, Fersht AR (2008) Algorithm for prediction of tumour suppressor p53 affinity for binding sites in DNA. *Nucleic Acids Res.* 36:1589–1598.

[132] Jordan JJ, et al. (2008) Noncanonical DNA motifs as transactivation targets by wild type and mutant p53. *PLoS Genet.* 4:e1000104.

[133] Fiucci G, et al. (year?) Siah-1b is a direct transcriptional target of p53: Identification of the functional p53 responsive element in the siah-1b promoter.

[134] Levy Y, Wolynes P, Onuchic J (2004) Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci. U S A* 101:511–516.

[135] Elenko MP, Szostak JW, van Oijen AM (2010) Single-molecule binding experiments on long time scales. *Rev. Sci. Instrum.* 81:083705.

[136] Doyle PS, Underhill PT (2005) in *Handbook of Materials Modeling*, ed Yip S (Springer Netherlands), pp 2619–2630.

[137] Mueller F, Wach P, McNally JG (2008) Evidence for a common mode of transcription factor interaction with chromatin as revealed by improved quantitative fluorescence recovery after photobleaching. *Biophys J* 94:3323–3339.

[138] Rajagopalan S, Huang F, Fersht AR (2011) Single-molecule characterization of oligomerization kinetics and equilibria of the tumor suppressor p53. *Nucleic Acids Res.* 39:2294–2303.

[139] Tokino T, et al. (1994) p53 tagged sites from human genomic DNA. *Hum. Mol. Genet.* 3:1537–1542.

[140] El-Deiry WS, Kern SE, Pietenpol JA, Kinzler KW, Vogelstein B (1992) Definition of a consensus binding site for p53. *Nat. Genet.* 1:45–49.

[141] Weinberg RL, Veprintsev DB, Bycroft M, Fersht AR (2005) Comparative binding of p53 to its promoter and DNA recognition elements. *Journal of Molecular Biology* 348:589 – 596.

[142] Kochaniak AB, et al. (2009) Proliferating cell nuclear antigen uses two distinct modes to move along DNA. *J Biol Chem* 284:17700–17710.

[143] Halford S, Marko J (2004) How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.* 32:3040–52.

[144] Harms G, Sonnleitner M, Schuetz G, Gruber H, Schmidt T (1999) Single-molecule anisotropy imaging. *Biophys. J.* 77:2864–2870.

[145] Lomholt MA, van den Broek B, Kalisch SMJ, Wuite GJL, Metzler R (2009) Facilitated diffusion with DNA coiling. *Proc. Natl. Acad. Sci. USA* 106:8204–8208.

[146] Ito A, et al. (2001) p300/CBP-mediated p53 acetylation is commonly induced by p53-activating agents and inhibited by MDM2. *EMBO J* 20:1331–1340.

[147] Gu W, Roeder RG (1997) Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. *Cell* 90:595–606.

[148] Friedler A, Veprintsev DB, Freund SM, von Glos KI, Fersht AR (2005) Modulation of binding of DNA to the C-terminal domain of p53 by acetylation. *Structure* 13:629 – 636.

[149] Ang HC, Joerger AC, Mayer S, Fersht AR (2006) Effects of common cancer mutations on stability and DNA binding of full-length p53 compared with isolated core domains. *J Biol Chem* 281:21934–21941.

[150] Mirny LA (2010) Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences* 107:22534–22539.

[151] Anderson J, Widom J (2000) Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. *Journal of Molecular Biology* 296:979 – 987.

[152] Santo ID, Causa F, Netti PA (2010) Subdiffusive molecular motion in nanochannels observed by fluorescence correlation spectroscopy. *Anal Chem* 82:997–1005.

[153] Record MT, Zhang W, Anderson CF (1998) Analysis of effects of salts and uncharged solutes on protein and nucleic acid equilibria and processes: a practical guide to recognizing and interpreting polyelectrolyte effects, hofmeister effects, and osmotic effects of salts. *Adv Protein Chem* 51:281–353.

[154] Levy Y, Onuchic JN, Wolynes PG (2007) Fly-casting in protein-DNA binding: Frustration between protein folding and electrostatics facilitates target recognition. *Journal of the American Chemical Society* 129:738–739 PMID: 17243791.

[155] Huang Y, Liu Z (2009) Kinetic advantage of intrinsically disordered proteins in coupled foldingbinding process: A critical assessment of the fly-casting mechanism. *Journal of Molecular Biology* 393:1143 – 1159.

[156] Chen J (2009) Intrinsically disordered p53 extreme C-terminus binds to S100B($\beta\beta$) through "fly-casting". *J. Am. Chem. Soc.* 131:2088–2089.

[157] Wilder PT, Rustandi RR, Drohat AC, Weber DJ (1998) S100B($\beta\beta$) inhibits the protein kinase C-dependent phosphorylation of a peptide derived from p53 in a Ca2+-dependent manner. *Protein Science* 7:794–798.

[158] Bellanger S, et al. (2011) Tumor suppressor or oncogene? A critical role of the human papillomavirus (HPV) E2 protein in cervical cancer progression. *Am. J. Cancer Res.* 1:373389.